# COMPUTED OBJECTIVITY

## A CRITICAL REFLECTION ON USING DIGITAL METHODS IN THE HUMANITIES

by

**MARTIJN WEGHORST**

A thesis submitted in partial fulfilment of the
requirements for the degree of

**Master of Arts in New Media and Digital Culture**

at the

**Universiteit Utrecht**

***2015***

# TABLE OF CONTENTS

# ABSTRACT

*In this explorative thesis, I will show how I used a dataset of 50 million tweets written during the 2014 World Cup to try to answer prevailing questions about the current media landscape. The focus, however, is not on the results of these case studies but on the process of getting to them. As I was continually confronted with making subjective decisions and with questions about the validity of results, I developed awareness of the fact that data and digital methods are never fully neutral. I extended Daston and Galison's concepts of mechanical objectivity and trained judgment to include digital methods, introducing computed objectivity and trained data judgment. Finally, I argue that it is crucial for media scholars to develop a critical stance towards big data research.*

# 1. INTRODUCTION

The 2010's can be regarded as the decade in which sports spectatorship truly expanded to the online world. Although Twitter and Facebook already existed during the 2008 Beijing Olympics, their user bases were still not nearly as sizable as they were during the 2010 World Cup and 2012 London Olympics[1] (Associated Press, 2012). Twitter revealed that the London Olympics opening ceremony, which was televised worldwide, alone generated more than 9.5 million tweets.[2]

This development coincides with a second wave of digital humanities, in which born-digital data is used as a source of study and to give insight into prevailing issues in the humanities. Because of my interest in both digital methods and sports, I decided early on to study a corpus of tweets related to major sports events. After I collected my data (first during the 2012 Olympics, later during the 2014 World Cup) I did not yet have an exact research question. Because of this uncertainty I simply started 'playing' with my data and found that it offered many possibilities, starting points and different ways of working with it.

[1] At the Summer Olympics in Beijing in 2008, Twitter had about 6 million users and Facebook 100 million. In 2012, the figure was 140 million for Twitter and 900 million for Facebook

[2] http://blog.uk.twitter.com/2012/07/today-on-twitter_28.html

### *Methodology*

Initially, I planned to use a dataset of 50 million tweets written during the 2014 World Cup to try to answer prevailing questions in media studies. Such a dataset may reveal something about media consumption and use, as we can assume people were simultaneously watching television and using Twitter during matches. I conducted three different case studies, all of which were focused on the current media landscape:

- Media production: what links are shared by accounts of media organisations;
- Media consumption: are online media still defined by geographical boundaries;
- Media use: how and why are people simultaneously watching television and using Twitter.

As my research progressed I quickly found that data-driven research (in the sense of big data, and not of interviews) is prone to unique dangers and perils, and I was increasingly confronted with relatively subjective decisions and with questions about the validity of certain results. I found that Twitter data is never truly raw, and any analysis of it is full of assumptions and black boxes.

I decided to focus my thesis on my development of an awareness of those dangers and assumptions instead of on the results of my initial research, relating it to Daston and Galison's concept of *trained judgment*. The emergence of digital methods in the humanities is a relatively recent development and although  am by no means arguing that such research is useless, I do think that media scholars need to develop literacy of, and a critical stance towards it. During every data-driven study personal assumptions and design decisions influence results, of which my own research will serve as an illustration. Scholars should never stop to critically reflect on those decisions. Finally, I aim to present an explorative study of and a critical reflection on the epistemic value of big data and digital methods for the humanities.

I will start my thesis by briefly defining what digital humanities entail and by identifying some of its key challenges. After that, I will elaborate on how I think digital methods fit in Daston and Galison's history of objectivity. Then I will go into the construction of my dataset and present three case studies I conducted with my dataset of 50 million tweets. I will illustrate my own judgment training and the pitfalls I encountered by giving detailed descriptions of each research decision I made.

# 2. DIGITAL HUMANITIES

> *Data […] form the bedrock of modern policy decisions by government and nongovernmental authorities, […] underlie the protocols of public health and medical practice, [...] inform what we know about the universe, and help indicate what is happening to the earth's climate.*
> (Gitelman and Jackson)

This notion of big data as the centre of knowledge is also spreading within the humanities and social sciences. The wish of these disciplines to be seen as 'objective' and to compete with natural sciences, paired with the availability of big datasets, has resulted in an increased interest in doing data-driven research (Latour 2009).

In the following paragraphs I will first explain how the widespread availability of computers changed research within the humanities and resulted in the formation of *digital humanities*. After that I will elaborate on Rogers' concept of natively digital data and methods and present some of its pitfalls and challenges. Finally, I will relate this to Daston and Galison's history of objectivity and show how important experience and an expert-eye can be while doing data-driven research, introducing the concepts of computed objectivity and trained data judgment.

## 2.1. Digitizing data

There are a lot of definitions of 'digital humanities', but I will proceed by using the one given in Digital Humanities Manifesto 2.0 (Schnapp and Presner 2009):

> *Digital Humanities is not a unified field but an array of convergent practices that explore a universe in which […] digital tools, techniques, and media have altered the production and dissemination of knowledge in the arts, human and social sciences.*

These practices started with translating existing data and methods to a digital format and capitalizing on the processing power of computers (Presner 2010, Schnapp and Presner 2009) and on the widespread availability of the World Wide Web. In *Virtual Ethnography*, Hine (2000) examines ways for ethnographers to utilize the Internet. In *Doing Internet Studies*, Jones (1999) presents some relatively early texts on doing sociological research on, and using, the Internet. Publicly available data was scarcer than it is today, and researchers were still developing feasible research methods. Witmer et al (1999) present a methodology for performing survey-based research on the

Internet, and in *Complementary Explorative Data Analysis* Sudweeks and Simoff (1999) provide a hybrid form of qualitative and quantitative research.

Besides digitizing methods, scholars have also digitised data, enabling them to quantitatively study societal and cultural change. Michel et al. (2010) studied “a corpus of digitised texts containing about 4% of all books ever printed” and found that it could provide insights into various fields, including the evolution of grammar and the adoption of technology. They also claim that their dataset can reveal something about society, as the use of the word slavery peaks during the Civil War and again during the civil rights movement, for example.

## 2.2. Natively digital data

As opposed to digitizing methods and data, known as the first wave of digital humanities, “the […] second wave [of digital humanities] is deeply generative, creating the environments and tools for producing, curating, and interacting with knowledge that is ‘born digital’ and lives in various digital contexts” (Presner 2010). Imagine browsing Facebook or doing a Google search: while it may seem that you are not explicitly posting new content to these websites, they do save a variety of information and automatically generated metadata. They track where you are located, the links you click on, how long a certain page is displayed, etcetera. Rogers (2013) calls this *natively digital data:*

> *An ontological difference may be made between the natively digital and the digitized, that is, between the objects, content, devices, and environments that are ‘born’ in the new medium and those that have ‘migrated’ to it.*

Rogers argues that this distinction not only applies to data itself, but also to the methods for studying data. He advocates that instead of transforming offline methods, we must try to develop natively digital methods for studying the web that are as close to the subject of research as possible. Furthermore, he proposes that we should study a phenomenon not just by itself, but also by looking at the data or by-products it generated. These by-products are often free of transformations and invisible to the end user, as opposed to surveys and questionnaires that are prone to *social desirability bias* (for example, young people might understate the amount of time they spent watching television or playing games). As a result, by-products are “more effective in understanding people’s true feelings and attitudes” (Harrington).

A recent post on Hacker News discussing Draftback illustrates this in one of the most extreme ways I have seen so far. Draftback is a browser extension that visualizes every keystroke that was ever made during the editing of a Google document. This enables the reader to look into the 'archaeology of writing', as creator James Somers calls it, and see how a writer has arrived at a certain passage. Clearly, this example shows the difference between digitised books, which are nothing but scanned words, and digitally written texts. Maybe even more importantly, it also demonstrates how data can be invisibly generated as an unmediated by-product of Internet use; it is quite possible the writer was never aware of the fact that the complete revision history of his document was being saved[3], and it is evident that it was never Google's intention to provide such an archaeology.

***Online-groundedness***

Rogers finally argues that natively digital data may move beyond the study of online culture only, introducing the term *online-groundedness*. He proposes "a research practice that learns from the methods of online devices, repurposes them and seeks to ground claims about cultural change and societal conditions in web data". In other words: we should view the web not just as an object of study, but as a source as well. An example that Rogers touches upon is detecting influenza pandemics. Google collects the frequency of the use of certain search terms related to the flu to determine when there is a peak and, thus, possibly an outbreak. The same has been tried using Twitter messages and despite the potential ambiguity of these messages, there is a "high degree of correlation between pre-diagnostic social media signals and diagnostic influenza case data" (Collier et al. 2011).

## 2.3. Issues and challenges

The availability of natively digital datasets clearly offers many opportunities. However, the assumed objectivity of big data also creates new issues and challenges that need to be addressed. Pilkey and Pilkey (2007) criticize quantitative research by explaining how it can be "biased, skewed or slanted by political correctness, advocacy, or economic interests". Borgman (2009) argues that "despite many investments and years of development, basic infrastructure for the digital humanities is still lacking". Therefore, Berry (2012) calls for a third wave of digital humanities:

> *[We should] look at the digital component of the digital humanities in the light of its medium specificity, as a way of thinking about how medial changes produce*

[3] In fact, there was quite some discussion when users found out this revision history was visible whenever a document was shared with other users.

> *epistemic changes. […] To understand the contemporary born-digital culture and the everyday practices that populate it – the focus of a digital humanities second wave – we need an additional third-wave focus on the computer code that is entangled with all aspects of culture and memory".*

### *Black boxing*

Rieder and Röhle (2012) present some major challenges for big data research, one of which is *black boxing*. While software and algorithms influence the results of data-driven research, their underlying assumptions and methods often remain unclear:

> *Many of the techniques issued, for example, from the field of machine learning show a capacity to produce outputs that are not only unanticipated but also very difficult for a human being to intellectually reconnect to the inputs.*
> Rieder and Röhle (2012)

In the rest of this thesis I will regard black boxing not only as the invisible technology at work in digital tools, but also as the collection of personal decisions and assumptions made by the scientist that can come to surface when writing or choosing codes and algorithms. Anything that affects the final results of a study, but is not explicitly visible in those results is part of the black box.

Another issue that Rieder and Röhle (2012) identify is that using digital methods constitutes a *Lure of objectivity*: "Questions of bias and subjectivity, which the computer was thought to do away with, enter anew on a less tangible plane - via specific modes of formalisation, the choice of algorithmic procedures, and means of presenting results." Furthermore, they mention the *Power of visual evidence*, arguing that "the use of numbers and images in scientific communication has a certain tendency to overwhelm the recipients, rendering them less prone to question their inherent explanations". In other words, numerical and visual results are often more easily accepted as evidence, because they are harder to reproduce or reverse to their negative argumentation.

Rieder and Röhle finally conclude "there needs to be a high degree of transparency regarding assumptions, choices, tools, and so forth". Borgman (2009) makes a similar argument:

> *Lacking an external perspective, humanities scholars need to be particularly attentive to unstated assumptions about their data, sources of evidence, and epistemology. We are only beginning to understand what constitute data in the*

*humanities, let alone how data differ from scholar to scholar and from author to reader.*

## 2.4. Trained judgment

The main issue of data-driven research is thus that it has such a strong persuasive quality, despite possibly being biased and the result of algorithms, software, and personal choices and assumptions. This can lead to a false sense of objectivity and for that reason we need to develop a critical understanding of such research and its epistemic impact.

Of course, academics have been confronted with issues of objectivity before: "Objectivity is situated and historically specific; it comes from somewhere and is the result of on-going changes to the conditions of inquiry" (Gitelman and Jackson 2014). Daston and Galison's (2007) groundbreaking volume *Objectivity* provides a history of the emergence of visual scientific objectivity by differentiating between three types of it:

- Truth-to-nature: portraying an "underlying type […], rather than any individual specimen";
- Mechanical objectivity: "an attempt to capture nature with as little human intervention as possible";
- Trained judgment: "[mixing] the output of sophisticated equipment with a 'subjective' smoothing of data".

Daston and Galison present these three types as consecutive steps in the evolvement of knowledge production, although not mutually exclusive. Rather, trained judgment supplements mechanical objectivity, as it entails that experts **interpret** mechanically produced images and deduce a (relatively) general truth from them. The importance of the process of creating scientific images is also discussed by Bredekamp et al. (2015), who claim that scientific images are "multilayered elements of the process of generating knowledge [that] play a constructive role in shaping the findings and insights they illustrate".

### *Computed objectivity*

What is so interesting about Daston and Galison's history is that it seems to have repeated itself. Data-driven research sometimes seems to be understood as a new form of mechanical objectivity that is independent of human intervention, creating a *computed objectivity*. What such an understanding forgets is that data are **never** entirely *raw*, as Gitelman and Jackson (2014) explain. A dataset is influenced by design decisions from early on in the process, like the choice of exactly what and when you will download, and what is allowed by the data source. During many steps, someone made a **choice**, and therefore we should not misconstrue data as being fully objective or neutral. Or as Manovich (1999) argues: "data does not just exist - it has to be generated".

The impossibility of fully eliminating human intervention in data-driven research is perhaps best explained by the analogy with photographic images given by Gitelman and Jackson (2014). When photography was believed by some to be fully mechanical and to do away with human intervention shortly after its introduction, others quickly pointed out that a photographer always *frames* what the camera captures. And as much as a camera does not position itself, an algorithm to collect or study data does not write itself; or as much as a photographer frames his subject, a programmer frames his data:

> *The presumptive objectivity of the photographic image, like the presumptive rawness of data, […] is not sufficient to the epistemic conditions that attend the uses and potential uses of photography. At the very least the photographic image is always framed, selected out of the profilmic experience in which the photographer stands, points, shoots. Data too need to be understood as framed and framing […].*
> Gitelman and Jackson (2014)

Moreover, the results of digital methods are prone to interpretation, despite their presumed sense of objectivity:

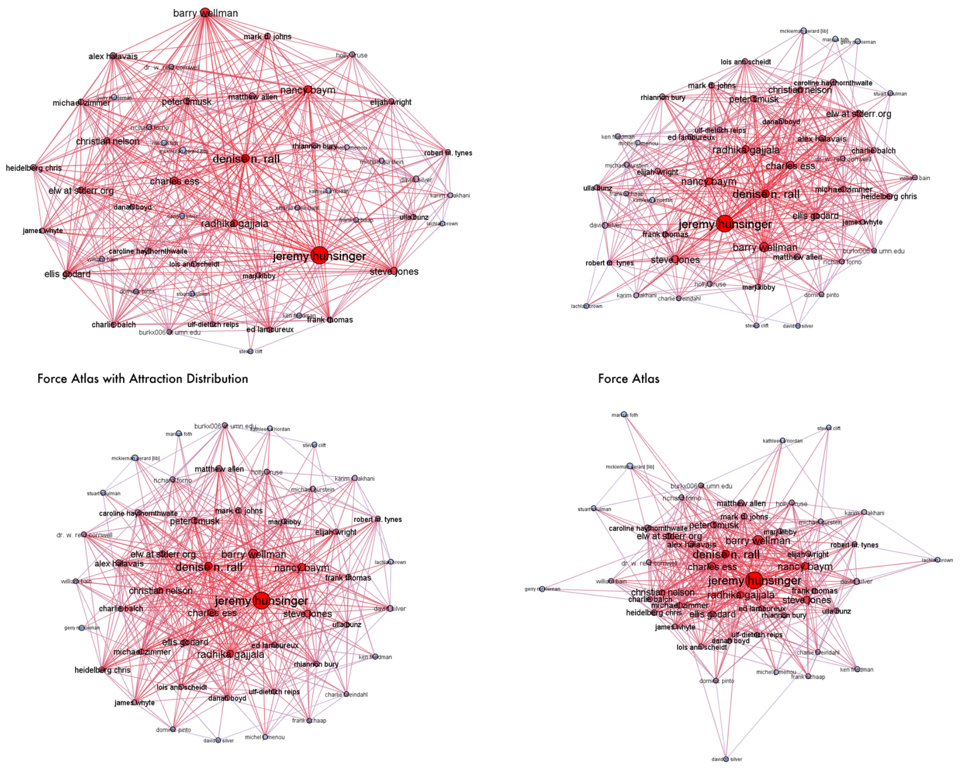
> *Interpretation is at the center of data analysis. Regardless of the size of a data, it is subject to limitation and bias. […] Data analysis is most effective when researchers take account of the complex methodological processes that underlie the analysis of that data.*
> boyd and Crawford (2012)

### *Trained data judgment*

Since data is framed and interpreted, and computed objectivity seems to reduce human intervention as little as mechanical objectivity, I argue that reflection on the epistemic value of digital methods is crucial and that scholars need to develop a *trained data judgment*. By doing actual data-driven research, while reflecting on decisions made and steps taken, experience and data literacy can be gained. Following that, media scholars can form a critical stance towards code and algorithms as scientific apparatuses, like they have done before for photography.

Reflection on the role and objectivity of data visualisations is not something new. Tufte famously criticised examples of (scientific) visual representations of quantitative information in *Envisioning Information* (1990) and *Visual Explanations* (1997), and Friendly (2005) presents the “historical developments leading to modern data visualization”. Rieder (2010), however, gives perhaps the best illustration of why developing a trained data judgment is important for humanities scholars. Similar to Daston and Galison’s (2007) illustration of trained judgment, who presented multiple images of the same galaxy “with the explicit aim of schooling the reader’s judgment”, he shows the different outcomes of different visualisation algorithms (figure 1).

The tool Rieder uses is Gephi, which can visualize networks of nodes and relations (Bastian et al. 2009). After importing a file you have to choose one of the available algorithms. However, these algorithms constitute a black box for most scholars, as the underlying calculations remain unclear. It is exactly this influence of code and algorithms that needs to be exposed in order to create awareness of the possible subjectivity, ambiguity, and bias of data-driven research.



Force Atlas with Attraction Distribution

Force Atlas

Fruchterman Reingold

Yifan Hu Proportional

**Figure 1 – The outcome of four different algorithms (Rieder 2010)**

### 2.5. Summary

Data-driven research can seem (mechanically) objective because results are the outcome of computers. The underlying processes and algorithms are very abstract and often remain invisible, and when they are visible they are often incomprehensible to humanities or media scholars. Because of this, the data-scientist may be viewed as a "will-abnegating worker" (Daston and Galison 2007) who merely executed these algorithms. However, the choice of which algorithms to use is never fully automated and relies on the experience and skills of the researcher. Furthermore, the researcher actively frames his data during the process of research, for example when he composes a corpus and writes or chooses algorithms.

Finally, I propose that it is crucial for media scholars to develop a trained data judgment. I will do so by giving both a critical reflection on and a factual report of the process of doing data-driven research, and by showing how inextricably linked these two things are. Research can benefit from a critical stance towards personal assumptions and choices as these matters can affect results, and forming a critical reflection is much easier when issues and challenges have been experienced first-hand.

# 3. TWITTER DATA AS OBJECT *AND* SOURCE OF STUDY

In this chapter, I will describe the difficulties that arose during the creation of my corpus of tweet data. I will first elaborate on the platform of Twitter itself and its system for distributing data and the general issues this brings forth. Then, I will give an account of the steps that need to be carried out by a researcher in order to construct a dataset, going into questions like which keywords to track and which data to include in the final sample.

My research started with a dataset of 50 million Twitter posts, or *tweets*. Twitter is a microblogging service where users can post 140-character messages, share interesting links or images, have conversations with and follow other users, and repost (*retweet*) or like other people's messages. The service can be used on any device and encourages people to share updates in real time. The default text that is displayed in the tweet-box is "what's happening", and on its homepage Twitter presents itself as a medium where you "Connect with your friends - and other fascinating people. Get in-the-moment updates on the things that interest you. And watch events unfold, in real time, from every angle.”

Because Twitter saves and publishes automatically generated metadata like the device or software the tweet was written with and the time zone of the user, we can study where, when and how people are ‘answering’ its question of what is happening:

> *The contributions of ordinary members of the public to microblogging services like Twitter can give social scientists unique insights into public reactions to specific events and to changes in public opinion over time. […] Microblogs are particularly useful for monitoring public opinion in this way because (a) they are reliably time-stamped, unlike most of the rest of the Web, so that they can be analysed from a temporal perspective, (b) they are relatively easy to create, so that a wide segment of the population with Internet access could, in theory, create them, and (c) they are public and hence accessible to researchers, unlike most social network sites.*
> Thelwall (2014)

Furthermore, the fact that users remain unaware of the fact that they might be the subjects of academic research makes Twitter data an excellent example of what Rogers describes as natively digital. The data is generated in an unmediated way and not collected through surveys or interviews:

> *Twitter […] provides us with exactly such a means of inconspicuously observing the activities of television audiences. We are left with [data] which are created ‘organically’, as the research participants are generally unaware of their own participation in a research study.*
> Harrington (2014)

A corpus of tweets qualifies as ‘good data’ as defined by Borgman (2009), since tweets can be delivered in high quantities that are all structured in the same clean way: “Good data are captured as cleanly as possible and as early as possible in their life cycle”.

Since so many people use Twitter during their daily activities, and since its usage data is so unobtrusively saved, it may tell us something about everyday life. Twitter may not just be an object of study but a source as well, delivering online data in which claims about the offline can be grounded. This makes the platform a very popular source of data for media scholars. For an extensive overview of doing research using Twitter data within humanities, see Weller et al. (2014).

A challenge of using Twitter data as a source of study is that Twitter use is not evenly distributed among people, as “age, gender, race/ethnicity, socioeconomic status, online experiences, and Internet skills all influence the social network sites people use” (Hargittai 2015). This raises the question of what Twitter-driven research represents. Perhaps it is only possible to ground claims about media use by people that **also** use Twitter.

Another thing to consider is the privacy of Twitter users. Although most users publish their tweets publicly, they may be unaware of all the metadata Twitter saves, like the device a tweet was posted with. Of even bigger concern is that leakages can occur when users do choose to post their tweets privately (Zimmer and Proferes, 2014). When a private tweet is retweeted by a user with a public account, or when a tweet was reposted to a public timeline before deleting it, the content is traceable. Furthermore, the streaming API does not give notices of tweet deletions, rendering it impossible to know which tweets are still public. For these reasons, I opted to hide all screen names from my research. This ensures that a tweet’s text and metadata cannot be traced back to its writer.

## 3.1 Twitter API

Twitter offers multiple *Application Programming Interfaces* (APIs) with which data can be collected.[4] One of these is the ‘Streaming API’, which pushes tweets to the client directly when they are posted. For a somewhat more detailed description of how I downloaded and saved Twitter data, see Appendix 1.

The data that the streaming API pushes include a tweet’s text, timestamp, location, retweet count, the device that was used, and any included images or URLs. In addition, it provides detailed information about the author, including the registration date, language, number of followers, website, and a short biography. Finally, it also includes status messages (*limit notices*) about the number of tweets that were omitted. This may happen due to sampling limitations, meaning that more tweets were being written per second than the streaming API connection could or may handle. When creating a timeline of your data, you have to take that number into account. In Appendix 2, I present a method for correcting rate limited streams. For a more extensive overview of data returned by Twitter’s APIs, see Bruns and Stieglitz (2014).

[4] Something to consider is that Twitter can pull the plug on any project that uses their API without giving further notice, as they did to Politwoops:
http://sunlightfoundation.com/blog/2015/06/04/eulogy-for-politwoops/

Any regular user account on Twitter is granted access to 1% of all tweets posted to the system (Gaffney and Puschmann 2014). Personally, I also never got more than 200 tweets per second, and others have reported limits of 50 tweets per second[5]. Gaffney and Puschman (ibid) claim that "while the [sampling] has never been independently verified as random, it is generally assumed that it is of an acceptable degree of randomness".

***Black box***

Not all data that the Twitter APIs return is fully neutral. For example, the language property that is returned is an automatic detection of which the underlying algorithm is unknown. Another example is the filtering of spam messages. I wanted to replicate the method of filtering spam as presented by Kwak et al. (2010). They deleted tweets containing one or more hyperlinks or trending topics and written by users that just joined. However, I found an indication that Twitter now spam-filters their streaming API, as the lowest account age I found in my database was 14406 seconds (the equivalent of 4 hours). I was not able to find Twitter's policy on spam in their streaming API anywhere. This is a clear example of black boxing, as we do not know the exact criteria of Twitter's spam filter and have to rely on our own findings.

Furthermore, data is influenced by the way it is entered into the Twitter platform. Users are encouraged to fill in certain information, like their time zone. But when doing so, Twitter offers a list of options that is alphabetically ordered, giving a time zone like Amsterdam an advantage over Zurich. This skews results, as people are likely to choose the first time zone corresponding to their own, instead of the exact location. In my dataset, Amsterdam (GMT+1) and Athens (GMT+2) were the most popular European time zones with around 2 million tweets, while Madrid (GMT+1) only had 500.000 and Helsinki (GMT+2) only 33.000, for example. This finding is a perfect example of the (necessity of) developing trained judgment, since it could only surface during actual research. It is not something that is intended by Twitter or something that is taken into account from the beginning; it **had** to be found during research at some point.

In conclusion, even the direct output of the Twitter API does not truly qualify as raw data. Twitter pre-filters and -processes the data that its API offers without disclosing criteria and scripts, and data can be ambiguous because of platform design. It is also worth noting that while I downloaded data directly from the Twitter API, other scholars

[5] https://dev.twitter.com/discussions/3914; https://dev.twitter.com/discussions/6349; https://dev.twitter.com/discussions/8923

may choose to use external tools like yourTwapperKeeper or 140kit. For an overview of tools, see Gaffney and Puschman (2014, p 60-64). When using an external tool more black boxing can occur, as the underlying scripts of the tool are unknown.

## 3.2. Constructing a dataset

One of the instances where I as a data scientist 'framed' my data was when I had to select the keywords I wanted track with my API connection. The amount of keywords that can be tracked using the streaming API was limited to 400 at the time, and exact matching of phrases was not supported[6]. Regardless of whether or not it would have been of interest, these API limits made it impossible to track all participating player's and coach's names. This and the fact that Twitter limits the amount of tweets that can be collected in a given timeframe certainly had an impact on some of my design decisions and shows that scrapers are not just objective tools, or as Marres and Weltevrede (2013, p. 316) put it: "Analysis precedes data collection, rather than succeeding it".

I decided to focus on keywords that either describe the complete event, or one of the matches in particular. I included descriptors for the whole event in multiple variations and languages ("world cup", "copa mundial"), and included two hashtags for each match: the ISO codes of the playing nations with and without "vs" as a separator ("#ESPNED" and "#ESPvsNED").

When querying a combination of words, tweets that contain all of them anywhere in the text (not necessarily in sequence) are included. For example, when tracking "Brazil 2014", tweets with hyperlinks to blogs or news articles written in 2014 and about Brazil are included[7]. In this case it was acceptable because the stream was active for only four weeks and the World Cup attracted most of the attention, but one could think of situations where this would not be desirable. Of course, it would be possible to post-process the data and filter out tweets where words are not occurring in sequence.

When composing a list of the most retweeted statuses in my database I stumbled upon a problem of *representativity*. Because my dataset was based on hashtags, not all tweets about the subject were included: "[the dataset] represents only a subset of all communicative activity which may be relevant to the themes described by the keywords or hashtags themselves" (Bruns and Stieglitz 2014). Because my dataset consists of tweets written during matches and a global television broadcast, including a hashtag to

[6] https://dev.twitter.com/streaming/overview/request-parameters#track

[7] For example: www.forbes.com/sites/kenrapoza/2014/06/16/three-scenarios-for-brazil-after-elections

clarify the subject was relatively unnecessary (ibid). A significant part of the relevant tweets may have simply mentioned a player’s name or the fact that a goal was scored, and thus has not been captured by the API. As a result, creating a list of the most popular statuses or most influential users is impossible. For more on the completeness of my dataset, see Appendix 2.

### *Data cleaning*

The format in which the Twitter API returns its data in is *JavaScript Object Notation* (JSON), a widely accepted open standard consisting of multidimensional key-value pairs that is used by almost all APIs. JSON-formatted data can be easily parsed with practically every programming language. To illustrate, a fragment of a tweet as returned by the Twitter API looks as follows:

```
"created_at": "Wed Jun 06 20:07:10 +0000 2012",
"id_str": 210462857140252672,
"text": "Example of JSON-formatted data",
"retweet_count": 66,
"user": {
  "screen_name": "twitterapi"
  "created_at": "Wed May 23 06:01:13 +0000 2007",
  "location": "San Francisco, CA"
}
```

When all was said and done and German player Mario Götze scored the deciding goal in the final match, my cloud server had collected hundreds of gigabytes worth of tweets and metadata. This forced me to clean my dataset, a subjective process:

> *The design decisions that determine what will be measured also stem from interpretation. For example, in the case of social media data, there is a ‘data cleaning’ process: making decisions about what attributes and variables will be counted, and which will be ignored.*
>
> boyd and Crawford (2012)

The first step of reducing the size of the dataset was to remove unnecessary fields like the *profile background colour* and to keep only information that could be relevant for my study. Unfortunately, this turned out to be a repetitive process as I threw away too much data on multiple occasions. Each time I realised that this had happened I had to re-parse the original files, which cost me hours of waiting. Therefore, I **strongly** advise any data researcher to play around with a subset of their dataset first before cleaning up the complete set, and to always save the original files.

This reduction still resulted in a corpus of which the size was beyond the scope of my technical abilities, so I decided to filter out more tweets. Twitter published some interesting facts about the World Cup-talk and stated that "of the 672 million total Tweets, we saw the bulk of the conversation during the live matches" (Rogers 2014). This gave me the idea to include only tweets that were explicitly about a match, and one match only. In other words: tweets that did not include a match hashtag or a hashflag of one of the participating teams, or included hashtags from different matches were excluded. Furthermore, I decided to include only tweets that were written between one hour before kick-off and one hour after the final whistle. I did this by using official liveblogs, and comparing the timestamp of each tweet to the start and end timestamps of the matches.[8] This resulted in a final dataset of 50 million tweets written **during matches** of the 2014 FIFA World Cup.

### *Trained judgment*

In conclusion, decisions must be made in order to be able to reduce or analyse a dataset. Raw data simply does not exist, and doing research with data means a sample has to be **constructed**. In this case I had to choose which keywords to track and decide how many and which data to include in the final sample. I made these choices and decisions based upon experience gained during previous research. I would not have made the decision to study millions of tweets had I not done research on smaller samples before. Similarly, the decision to track match hashtags would have been hard to make without prior knowledge, simply because these hashtags were starting to be used only shortly before matches. In fact, after the tournament I learned that it might have been better to include match hashtags with the participants in both sequences (#ESPvsNED and #NEDvsESP).

Being aware of and able to competently engage with these challenges and decisions is exactly what trained data judgment entails. Much like the subject of a photo has to be framed, data has to be constructed, and this is something that should happen consciously and preferably based upon experience.

[8] A dataset obtained through the Twitter streaming API can include tweets that were posted before you started collecting data. These are tweets that were retweeted later. So, even when a connection with the Twitter API is only open during the desired timeframe, you still need to check if a retweeted status meets the requirements.

### 3.3. Summary

In this chapter, I introduced Twitter as a source of study. I have shown that the output of the Twitter API is not truly ‘raw’, and that data gets ‘cooked’ during the post-processing and cleaning of that output. Because of API limitations, a number of tweets may be omitted during popular events. Sometimes, the data that **is** put through is the result of a black box, as Twitter returns an automatically detected language of a tweet and seems to apply a spam-filter.

Furthermore, choices have to be made about which keywords to track. The output of the Twitter streaming API sometimes needs to be cleaned before being useable or queryable, which can be a subjective process. Like a photographer that frames what is in front of his camera, a data researcher frames what gets passed through his scripts or software. Often, a researcher has a clear research question and a hypothesis *before* collecting and cleaning data, and this might influence the process.

# 4. USING TWITTER DATA ANALYSIS IN MEDIA STUDIES: CASE STUDIES

In order to demonstrate the types of research that can be conducted using digital methods I extracted three case studies from my initial research, which I will present in ascending order of complexity. The case studies are meant as an instructive guideline for doing research with natively digital data and as an illustration of how trained judgment can be developed, as I will describe how the process of research led me to question my own methods and assumptions. Moreover, each case study is related to a prevailing question within media studies and describes a particular challenge or opportunity. Consequently, I will demonstrate that (born-digital) data-driven research can be a useful addition to media studies.

The first case study goes into the production side of media and pertains to the links that are shared by media accounts. It is relevant for its display of (a lack of) representativity of a set of tweets. When tracking a subject using certain hashtags, not all content related to that subject is gathered. The second case study aims to answer whether media are primarily read globally or domestically. During this study I developed a method for determining the location of Twitter users, possibly creating a black box. The third and final study is about media meshing, or how people are simultaneously watching television and using Twitter. In this study I show how a table of figures can constitute a false sense of objectivity, and how a hybrid form of looking at aggregates and individual data points may provide insights.

## 4.1. CASE STUDY 1: MEDIA CONVERGENCE

In this first case study I will show issues related to filtering data based on hashtags. The study goes into the production side of media and the links and retweets shared by media accounts. Giving insight into the use of Twitter by (traditional) media could possibly provide a different perspective on studying media convergence, as it may reveal that companies indeed use Twitter as an additional node in their flow of content:

> *By convergence, I mean the flow of content across multiple media platforms. […] In the world of media convergence […] every consumer gets courted across multiple media platforms. Right now, convergence culture is […] shaped by the desires of media conglomerates to expand their empires across multiple platforms and by the desires of consumers to have the media they want where they want it, when they want it, and in the format they want…*
> Jenkins (2006)

Results of a survey research by Neuberger et al. (2014) have already shown that 97% of media organisations use Twitter to "attract readership" and to "advertise their own Internet content". I tried to do similar research using my sample of tweets and found that during World Cup matches around 99% of the links shared by media organisations indeed directed to their own website or social media accounts. As discussed before, however, this thesis is not concerned with making bold claims such as the above. Rather, I would like to explain the underlying assumptions and choices that influenced the resulting figure. For the actual study, see Appendix 3.

The first issue with this study is that it is inherently influenced by which keywords and hashtags were tracked. Besides the more general issues of representativity that I described before, there is another issue at stake here: the media organisations may have chosen to include hashtags more often when posting links to their own website, because "hashtags […] represent a convention designed to make tweets more easily discoverable" (Bruns and Stieglitz 2014). Perhaps media employ a script that automatically fetches their liveblog, adds a hashtag and link to each new entry, and posts it to their Twitter account. Additionally, it is possible that media **do** link to external websites, but do so without including hashtags as these links exist outside of their automated feed.

Since Twitter does not disclose what accounts are owned by media organisations, I had to come up with a method to identify media accounts. I composed a set of requirements, including the minimum account age and number of followers, and created a sub-dataset with accounts that met all requirements. After that I further reduced the subset by looking at descriptions and screen names and handpicking accounts that qualified as media organisations. Of course, this shows how personal knowledge and decisions may have influenced the results: non-Western media organisations may be excluded simply because I did not know them.

In conclusion, even a relatively simple analysis of shared links by a group of accounts is prone to personal assumptions, knowledge, and design decisions. While the resulting numbers certainly indicate a significant degree of media convergence they need to be read with great caution, as they may be a consequence of inconsistent hashtag use or only apply to well-known Western media, since these were picked unevenly often.

## 4.2. CASE STUDY 2: MEDIA GLOBALISATION

In this second case study I will show how I formalised a method to detect geographic information for tweets and how I tested its reliability. The question I studied was whether media websites are read by an international audience. I believe my dataset is particularly suitable for this purpose as we can assume that people all over the world are interested in its subject and that most national news media produce coverage of it.[9] The results of the study are presented in Appendix 4.

Although Twitter provides some geographical data, “location sparsity remains a commonly remarked-upon research issue” (Wilken 2014). Only a fraction of tweets comes with exact geographical coordinates, because this is a setting that has to be explicitly turned on by users. In my dataset only 3,65% of tweets (1.935.461) contained coordinates. Furthermore, while doing this research I found that the Twitter API does not disclose a user’s coordinates when he simply hits the retweet button. Since this is not something that is made clear by the Twitter API documentation, it shows how crucial experience or trained judgment is for data scientists.

[9] For this reason, other datasets may give other results. When studying a subject like national politics or Major League Baseball, results could be different either because only Americans are interested, or because only American media cover it (sufficiently).

### *Geoparsing*

Because I needed a bigger sample of geo-tagged tweets and, more importantly, one containing retweets, I was forced to develop a custom method for detecting a user's location. One of the alternative methods for detecting geographical locations is *geoparsing*, a method that uses (parses) textual content to estimate a location. Because such content does not follow strict conventions in most cases, geoparsing usually involves digital gazetteers, or "place name directories containing names, spatial references, feature types and additional information for named geographic places" (Janowicz and Keßler 2008, p. 1129, as cited in Wilken 2014).

I chose to automatically identify the geographic location of tweets by parsing the user location text and match on country names. This text can be set to anything the user chooses. I used a list with country names in both English and in the language(s) spoken in that country (for example Schweiz, Suisse and Svizzera as alternatives for Switzerland). I excluded the Maldives and Sri Lanka because of difficulties with their script, and the Congo's because of possible ambiguity. After parsing, I removed tweets where the user's location contained more than one country name. Although on the one hand one could argue this creates a new layer in the black box, others argue that developing new methods may lead to a wider toolkit for science in general:

> *Additional, more specific metrics which relate to particular communicative contexts on Twitter may also be developed; we encourage researchers to document their analytical choices in such specific cases in similar detail, so that these metrics can also become part of the wider toolkit of conceptual models and practical methods which is available to social media researchers.*
> Bruns and Stieglitz (2014)

See Appendix 7 for a full list of country names, both in English and the languages spoken in the respective countries, and the script that parses user biographies.

### *Reliability*

The scripted detection resulted in a corpus of 11.552.082 tweets (of which 45% were retweets) for which I had identified the location of the author – nearly six times as big as the number of tweets with coordinates. I checked the reliability of my method by comparing these tweets to the ones with explicit coordinates and found that 91,1% matched. One of the causes of non-matches is that location texts can be ambiguous: Armenia, for example, is not only the name of a country but also of a city in Colombia. Other causes are that people are only visiting Brazil temporarily, and that organisational accounts dedicated to the World Cup locate themselves in Brazil regardless of the real location of their writer/owner. This reveals an unanticipated advantage of my method: it measures nationality, whereas Twitter returns a user's current location.

One of the issues of identifying countries from a user's location text is that its success rate is not evenly distributed over countries.[10] American citizens, for example, set their user location to a combination of city name and state abbreviation.[11] This causes comparisons of absolute numbers between countries to be invalid and renders claims about the geographic distribution of particular accounts or websites useless. Instead, rankings of the most frequently retweeted accounts and shared domains can be made, per country. Using all those lists, I could see the ranking in each country per account, giving an indication of the popularity of any medium in any given country. The rankings are based on unique users and only countries with at least 500 different users are included. Such a clarification could easily be left out, although it can affect results heavily as a certain user could, for example, post 100 links to a website.

Finally, it is worth northing that people who have their country name in their biography **might** be more nationalistic and more likely to read domestic media over foreign media, skewing results. This again displays how important trained judgment or awareness of external factors is while doing such research: there can be many uncertain factors that influence results without being visible within the dataset.

***Summary***

I have presented a newly developed method for identifying Twitter user locations. By doing so, I have aimed to show that not every addition or personal design decision necessarily leads to a new layer in the black box. When giving a detailed description of the process and requirements, as well as possibly publishing source code, a new metric may become part of a wider toolkit of methods available to media scholars. At the same time, I have shown how newly developed methods can raise new issues, in this case the uneven distribution among countries. Such a reflection is precisely what trained judgment within digital humanities entails: since methods are often new, critically assessing them is pertinent.

[10] This is also one the reasons for choosing not to use friendship/follower data. Although downloading a sample of ESPN or BBC followers would have been possible it could not have led to claims about the geographic distribution of those followers. Alternatively, I could have downloaded the relations of a sample of users from a few countries and study what percentages follow a certain account. However, that would have cost too many time and API calls.

[11] I thought about using the names of a few big cities to determine user locations, but realised that those cities often have a relatively high number of foreigners in comparison to other parts of a country.

## 4.3. CASE STUDY 3: MEDIA MESHING

In my third and final case study I looked into media meshing, or the way people simultaneously watch television and use Twitter. We can assume that a significant part of Twitter users were simultaneously watching a match and discussing it online, as indicated by Fraser (2010): “the only real-time #WorldCup global viewing party will be on Twitter”. I aimed to find out *how* and *why* people post something to Twitter during a football match. Are there differences in percentages of retweets, replies, shared images or mobile use when comparing between matches? Are people reporting results or are they merely expressing emotion for personal satisfaction? For a more detailed overview of the answers to these questions, see Appendix 5.

In the beginning I did everything on-the-flow using tools like PhpMyAdmin and Excel, but this resulted in a pile of temporary spreadsheets, texts, and tables. At one point I wanted to compare earlier analyses of Dutch matches to German ones and realised I had forgotten half of my methods, causing me to redo the first analysis. I found that I was working rather impulsively and switching between chapters each day. When returning to a chapter after a while I had forgotten parts of my method and was unable to quickly reproduce the results. This caused me a lot of extra work when writing my final texts.

Because of this I decided to build a custom dashboard for accessing the data, as I would need to code all queries and analysis needed for displaying results. My way of trying to find answers to questions about Twitter use was to create a dashboard with as many different calculations in it as possible, and then pick out those irregularities that caught my attention. Such a dashboard can also have the advantage of being publicly accessible. It would be preferable to publish the raw data as well, but I decided not to due to its size.

### Tweet characteristics

First, I linked tweets to the corresponding moments in matches on a per-minute basis and showed *when* people are using Twitter during matches, similar to Rios’ (2014) article about penalty shootouts. In support of this I needed factual data about the matches containing timestamps and important match events (goals, red cards, first and final whistle). I chose to use the most primary source available, and obtained the data by downloading all 64 liveblogs from the official FIFA website (the organiser of the tournament).

Because studying the plain number of tweets per minute is not indicative of the motivations for using Twitter, I identified various tweet characteristics. I distinguished retweets and conversations by looking at whether the tweet text started with "RT @“ or “@”, respectively. I categorised tweets in mobile vs. non-mobile and in iOS and Android groups by looking at all distinct values of the ‘source’ field and manually determining their hardware and software. I separated tweets with links, images and mentions by looking at the ‘entities’ fields. I excluded retweets from the mentions.

### *Visual evidence*

For a few of these tweet characteristics, like the percentage of retweets and mobile/iOS/Android use, mentions, replies, images and links, I created graphs showing the progress of their values per match, with important match events annotated on the bottom. To be able to evaluate whether phenomena occurred during the whole tournament and not just during individual matches, I also calculated averages per match part. Finally, I summarised all my findings by putting them in a single table that displays each characteristic and its percentage per match part:

| | Pre | 1st half | Pause | 2nd half | Overtime | Penalties | Post |
|---|---|---|---|---|---|---|---|
| **Android** | 28.9% | 32.1% | 31.8% | 31.2% | 35.1% | 33.8% | 33.3% |
| **Replies** | 2.1% | 1.7% | 2.0% | 1.5% | 1.8% | 0.83% | 1.7% |
| **Avg. length** | 95 | 81 | 95 | 83 | 81 | 57 | 102 |
| **Foursquare** | 0.032% | 0.020% | 0.011% | 0.006% | 0.004% | 0.002% | 0.002% |
| **Images** | 32.3% | 21.0% | 33.8% | 18.7% | 17.5% | 4.8% | 35.6% |
| **Fans** | 6.1% | 5.0% | 6.8% | 5.2% | 4.8% | 2.2% | 7.3% |
| **Instagram** | 0.40% | 0.25% | 0.30% | 0.17% | 0.14% | 0.037% | 0.29% |
| **iOS** | 33.4% | 34.4% | 33.3% | 33.6% | 37.2% | 40.0% | 33.6% |
| **Links** | 9.1% | 11.0% | 15.6% | 11.6% | 9.0% | 9.5% | 15.6% |
| **Mentions** | 17.2% | 13.0% | 17.2% | 12.7% | 11.0% | 5.6% | 19.5% |
| **Mobile** | 72.7% | 74.4% | 75.9% | 73.3% | 77.9% | 74.4% | 75.4% |
| **Negative** | 0.35% | 0.66% | 0.66% | 0.92% | 1.2% | 1.1% | 1.0% |
| **Positive** | 1.5% | 1.4% | 1.2% | 1.3% | 1.2% | 0.88% | 1.6% |
| **Retweets** | 54.1% | 46.9% | 63.8% | 46.4% | 47.3% | 33.9% | 67.4% |

**Table 1 – Tweet characteristics over the course of a match**

While certainly showing some interesting numbers, this table was the result of so many assumptions, calculations and choices that it is practically impossible to trace back to the raw data and recreate the exact same results. Thus, it constitutes what Rieder and Röhle call a lure of visual evidence and a black box.

For example, the difference between absolute versus relative numbers can be misleading. Upon first look, one could take away from this that fans of the participating teams are using Twitter mostly when the game is not in progress, as the 'Fans' figure drops when the ball is in play. However, the table displays the *percentage* of tweets written by fans. In reality, both fans and neutral viewers tweet more during the match. The alternative and more likely explanation is that neutral viewers have a higher threshold of tweeting about a match: they only do so when something important happens.

***Ambiguity***

The Twitter API does not directly return whether a tweet was written on a mobile device or not, but returns the name of the software that was used. To compose a list of mobile tweets I had to identify what software qualifies as mobile; for example "Twitter for Android". Some of the most-used devices are shown in Appendix 6.

I found that almost all of the matches in the knockout stage had a relatively high percentage of mobile use. This led me to calculate the percentages of mobile use in both the group and knockout stages: 74,8% and 77,6%, respectively. This *could* have something to do with people watching in bars more often when the tournament progresses, thus grounding a sociological claim like that in web data. However, it could also have something to do with which countries progressed, since some countries have a much higher percentage of mobile users than others. Again, this shows that results are prone to ambiguity and need to be very cautiously interpreted.

## Individual data points

I quickly found out that table 1 gave me starting points for further research rather than concrete results. It was impossible to draw conclusions from the numbers alone, and I needed to dive back into the data to find underlying causes for irregularities. Manovich advocates a similar mix of quantitative and qualitative research, where the former reveals patterns that are made sense of by the latter:

> *We can use computers to quickly explore massive visual data sets and then select the objects for closer manual analysis. While computer-assisted examination of massive cultural data sets typically reveals new patterns in this data which even best manual "close reading" would miss […] a human is still needed to make sense of these patterns. […] Ideally, we want to combine human ability to understand and interpret - which computers can't completely match yet - and*

> *computers' ability to analyze massive data sets using algorithms we create.*
> Manovich (2011)

This is exactly the same as what Latour describes in *Tarde's idea of quantification*: "the notion of navigation where we are able to [...] navigate on our screens from the individual data points to the aggregates and back". I have endeavoured to allow for at least some degree of navigation between patterns and their individual data points. Table 1 can be navigated stepwise[12]: first, it can be split so that variables are visible per match, per match-part. Furthermore, a graph of each match is generated in which variables are shown on a per-minute basis (see Appendix 8 for a graph of the percentage of mobile use during Spain versus The Netherlands). Finally, each match page also contains a search box wherein tweets about that specific match can be queried. Ideally, this search functionality could be narrowed down so as to be able to zoom in on specific data points in the graph above it, but such a project would exceed the scope of this thesis.

In a sense, this counteracts the issue of black boxing, as you get to go back to individual and real data points, rather than looking at aggregates only. For me, this is a crucial part of trained judgment: do not blindly trust numbers as results, but confirm findings by taking a closer look at individual data points. I will now present an example of finding irregularities that I addressed more closely.

### *Instagram links*

Two rows in table 1 that caught my attention were the percentages of tweets with images versus tweets with Instagram links. Although the differences seem small (relatively speaking, of course the dataset contains much more Twitter images), there is one important distinction. Twitter images are shared approximately the same number of times during pre-game, half time, and post-game discussion. Instagram links, on the other hand, are shared 30% more often during pre-game talk than during half time or after the game.

I did an assessment of 40 post-match images on both platforms (using England versus Italy and Brazil versus Germany) and this delivered the following results: of the 40 Instagram links, 24 directed to match photos or collages or memes[13] and 16 linked to

[12] See http://thesis.martijnweghorst.com

[13] Memes are online inside jokes, spread rapidly by Internet users and often with slight variations.

personal photos. Of the 40 Twitter photos none could be defined as a selfie[14] and only two were personal photos: one of a television and the other somewhere on a street. All of the others were either screenshots, Photoshopped images, match results graphics, or match photos.

What this closer look at individual data points thus revealed is that the types of images that are shared on Instagram and Twitter are very different. Instagram is a platform where people share relatively many photos of themselves and/or made by themselves, whereas on Twitter people mostly share memes and screenshots.

# 5. IMPLICATIONS FOR MEDIA STUDIES

In the previous chapter I have given various examples of challenges and issues that surfaced during my study of the current media landscape on Twitter. I started by showing how a data sample is never truly raw, and gets cooked already during its collection and cleaning process.

In the first case study, about Twitter use by media organisations, I showed that the fact that a data sample is based on certain keywords or hashtags can strongly influence the results. I also described how my personal knowledge and background could have affected the composition of a list of media organisations, despite setting strict requirements.

In the second case study I developed a new method for identifying a Twitter user's location and tested its reliability. By giving a detailed description of the process and by publishing the source code, I aim to show that such a custom development may lead to a wider toolkit of available metrics and methods.

In the final case study I presented a table with the results of a study of Twitter use during different parts of a football match. I linked tweets to match data provided by FIFA.com and showed what percentages of tweets contained images or links, were retweets, or were posted with a mobile device. I explained the ways in which this table

[14] A selfie is a photograph that one takes of him- or herself.

is a black box and has an aura of objectivity, although much of its contents were at least partly the result of my personal assumptions, decisions and knowledge. The figures on mobile use, for example, are a result of manually classifying the names of different software. Furthermore, I showed that the results can be interpreted in various ways and that the data does not simply speak for itself. Finally, I presented an example of the complementary qualities of pattern recognition and close reading. My table revealed an irregularity of images shared through Twitter and Instagram, and by doing a qualitative case study I was able to find differences in the uses of both media.

## 5.1. Key elements of trained data judgment

My personal process of doing (big-)data-driven research has led me to believe that the development of a trained judgment for such research among media scholars is an absolute necessity. Scholars must be trained in their judgment of underlying assumptions, structures and choices of data visualisations and the construction of data sets. Not just when they aspire an academic career but at all times, because there is also a more public demand of 'data literacy' since data visualizations increasingly appear in corporate and public administration decision making (Gitelman and Jackson 2014 p. 1, Bollier 2010).

This brings me to a list of some of the key elements of trained data judgment, which needs to be developed further in the future:

1. Be aware of the flaws and shortcomings of a dataset at all times. A dataset that is truly raw and 'objective' is impossible to find. Furthermore, keep in mind how these flaws affect the claims you can make and the conclusions you can draw.
2. Be aware of your personal background and assumptions.
3. When there is a need to develop new methods or metrics, try to describe them with as much detail as possible and keep in mind that others may want to replicate the method.
4. Often, a hybrid form of quantitative and qualitative research produces the most interesting results. Where big data analysis can reveal patterns and irregularities, looking at individual data points may reveal their underlying causes. Moreover, returning to individual data points restores a bit of context.

Each of the above points surfaced during my own research, which brings me to the final and perhaps most important element of trained data judgment: **always keep reflecting** on what you are doing. As I have aimed to show with my case studies, trained data judgment is something that must be developed and improved while doing actual research. For example, when you force yourself to document methods and algorithms and to outline every choice and decision made, new insights and issues with your data or methods might emerge.

# 6. COMPUTED OBJECTIVITY

We are currently seeing an abundance of publicly available data. One of these data suppliers is Twitter, a social media service on which users post short messages. As Rogers (2013) points out, these data can be used not just as objects of study, but as sources as well. This is a relatively new phenomenon within the humanities and media studies, and methods to deal with this data are still being developed.

With this thesis, I have tried to show how a media scholar can go from collecting a dataset to performing academic research, and how studying new data sources can enrich existing humanities research projects. In the first case study, for example, I quantitatively studied media convergence by looking at the links media organisations shared. And in the last of my case studies, I took a closer look at media meshing, or how people simultaneously use and consume multiple media. Whereas traditional research of this phenomenon is based on surveys, I used Twitter data to gain more insight into the different uses of Twitter during different moments in a football match.

During this research I became increasingly aware of the dangers and perils of using Twitter data as a source of research and came to the conclusion that using digital methods requires a new step of objectivity. Similar to Daston and Galison's description of mechanical objectivity, in which human intervention is eliminated as much as possible, digital methods seem to constitute a computed objectivity. Data are seen as neutral, and deploying algorithms appears to be strongly objective. In this thesis, however, I have aimed to show both are untrue, as data and algorithms are the result of design decisions, limitations and personal assumptions.

As Gelernter (2010) argues, we should not blindly trust, and always critically look at, the results of (big-)data-driven research. However, this does not mean that the digital humanities and data-driven research in media studies are pointless. Again similar to the development as described by Daston and Galison, we need trained judgment, or trained *data* judgment. By doing actual data-driven research, while reflecting on decisions made and steps taken, experience and literacy can be gained. Following that, media scholars can form a critical stance towards code and algorithms as scientific apparatuses, like they have done before for photography.

It has been my aim to convey my own awareness or trained data judgment to the readers of this thesis and other media scholars. As I shifted the focus of my thesis to the dangers and perils of digital methods, I transformed my case studies to illustrations of the ways that assumptions, decisions and limitations can influence results. I have shown the steps of the formation of a corpus of tweets, and have demonstrated that analysis starts before a single tweet is obtained, because of the choice of keywords to track. Furthermore, I have given an example of coming up with a new method when I needed to determine user locations. In the last case study I demonstrated the ambiguity and deceptive qualities of a table of plain numbers.

I argue that it is crucial for media scholars interested in data-driven research to be aware of their personal assumptions and decisions, and about technical limitations, as these can be of great influence on results. Reflection is essential for the development of that awareness, or trained judgment, as I have experienced myself and of which this thesis is a result. Finally, I have aimed to show how ways of studying data become ways of knowing, similar to how Daston and Galison have shown "how ways of seeing become ways of knowing".

# ACKNOWLEDGMENTS

I would like to take this opportunity to thank Mirko Schäfer for all his help during the writing of this thesis. Without him, I would have been a lot less content with the result.

# REFERENCES

Associated Press. (2012, June 20). London Games to be first social media Olympics. Retrieved from http://gadgets.ndtv.com/social-networking/news/london-games-to-be-first-social-media-olympics-233765

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In *Proceedings of 3rd International AAAI Conference on Weblogs and Social Media*.

Berry, D. M. (2012). Introduction: Understanding the Digital Humanities. In D. M. Berry (Ed.), *Understanding Digital Humanities* (pp. 1–20). Hampshire: Palgrave Macmillan.

Bollier, D. (2010). The Promise and Peril of Big Data (pp. 1–3). Washington, DC: The Aspen Institute.

Borgman, C. (2009). The Digital Future is Now : A Call to Action for the Humanities. *Digital Humanities Quarterly*, *4*(1). Retrieved from http://works.bepress.com/borgman/233

boyd, D., & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679. http://doi.org/10.1080/1369118X.2012.678878

Bredekamp, H., Dünkel, V., & Schneider, B. (Eds.). (2015). *The Technical Image: A History of Styles in Scientific Imagery*. Chicago: The University of Chicago Press.

Bruns, A., & Stieglitz, S. (2014). Metrics for Understanding Communication on Twitter. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.), *Twitter and Society* (pp. 69–82). New York: Peter Lang Publishing, Inc.

Collier, N., Son, N. T., & Nguyen, N. M. (2011). OMG U got flu? Analysis of shared health messages for bio-surveillance. *Journal of Biomedical Semantics*, *2(Suppl 5)*(S9), 18–26. http://doi.org/10.1186/2041-1480-2-S5-S9

Daston, L., & Galison, P. (2007). *Objectivity*. New York: Zone Books.

Fraser, L., & Sports Partnerships. (2014). Follow the #WorldCup action on Twitter. Retrieved May 7, 2015, from https://blog.twitter.com/2014/follow-the-worldcup-action-on-twitter

Friendly, M. (2005). Milestones in the History of Data Visualization: A Case Study in Statistical Historiography. In C. Weihs & W. Gaul (Eds.), *Classification: The Ubiquitous Challenge* (pp. 34–52). New York: Springer. http://doi.org/10.1007/3-540-28084-7_4

Gaffney, D., & Puschmann, C. (2014). Data Collection on Twitter. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.), *Twitter and Society* (pp. 55–68). New York: Peter Lang Publishing, Inc.

Gelernter, D. (2010, April 26). Gefahren der Softwaregläubigkeit. Die Aschewolke aus Antiwissen. *Frankfurter Allgemeine Zeitung*. Retrieved from http://www.faz.net/aktuell/feuilleton/debatten/digitales-denken/gefahren-der-softwareglaeubigkeit-die-aschewolke-aus-antiwissen-1606375.html

Gitelman, L., & Jackson, V. (2013). Introduction. In L. Gitelman (Ed.), *“Raw data” is an oxymoron* (pp. 1–14). Cambridge, MA: The MIT Press. http://doi.org/10.1080/1369118X.2014.920042

Hargittai, E. (2015). Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 63–76. http://doi.org/10.1177/0002716215570866

Harrington, S. (2014). Tweeting about the Telly: Live TV, Audiences, and Social Media. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.), *Twitter and Society* (pp. 237–248). New York: Peter Lang Publishing, Inc.

Hine, C. (2000). *Virtual Etnography*. London: SAGE Publications Ltd.

Janowicz, K., & Keßler, C. (2008). The Role of Ontology in Improving Gazetteer Interaction. *International Journal of Geographical Information Science*, *22*(10), 1129–1157.

Jenkins, H. (2006). Welcome to Convergence Culture. Retrieved from http://henryjenkins.org/2006/06/welcome_to_convergence_culture.html

Jones, S. (Ed.). (1999). *Doing Internet Research: Critical Issues and Methods for Examining the Net*. Thousand Oaks, CA: SAGE Publications Ltd. http://doi.org/doi: http://dx.doi.org/10.4135/9781452231471

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a Social Network or a News Media? *Proceedings of the 19th International Conference on World Wide Web*, 591–600. Retrieved from http://dl.acm.org/citation.cfm?id=1772751

Latour, B. (2009). Tarde's idea of quantification. In M. Candea (Ed.), *The Social after Gabriel Tarde: Debates and Assessments* (pp. 145–162). London: Routledge.

Manovich, L. (1999). Database as Symbolic Form. *Convergence*, *5*(2), 80–99. http://doi.org/10.1177/135485659900500206

Manovich, L. (2012). Trending: The Promises and the Challenges of Big Social Data. In M. K. Gold (Ed.), *Debates in the Digital Humanities* (pp. 460–475). Minneapolis: University of Minnesota Press.

Marres, N., & Weltevrede, E. (2013). Scraping the Social? *Journal of Cultural Economy*, *6*(3), 313–335. http://doi.org/10.1080/17530350.2013.772070

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., … Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(176). http://doi.org/10.1126/science.1199644

Neuberger, C., Hofe, H. J. vom, & Nuernbergk, C. (2014). The Use of Twitter by Professional Journalists: Results of a Newsroom Survey in Germany. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.), *Twitter and Society* (pp. 345–358). New York: Peter Lang Publishing, Inc.

Pilkey, O. H., & Pilkey-Jarvis, L. (2007). Mathematical Models: Escaping from Reality. In *Useless Arithmetic. Why Environmental Scientists Can't Predict the Future* (pp. 22–44). New York: Columbia University Press.

Presner, T. (2010). Digital Humanities 2.0: A Report on Knowledge. *Connexions Project*. Retrieved from http://cnx.org/content/m34246/1.6/

Rieder, B. (2010). One network and four algorithms. Retrieved from http://thepoliticsofsystems.net/2010/10/one-network-and-four-algorithms/

Rieder, B., & Röhle, T. (2012). Digital methods: Five challenges. In D. M. Berry (Ed.), *Understanding Digital Humanities* (pp. 67–84). Hampshire: Palgrave Macmillan.

Rios, M. (2014). Penalty kicks, as seen through Twitter data. Retrieved from https://blog.twitter.com/2014/penalty-kicks-as-seen-through-twitter-data

Rogers, R. (2013). *Digital Methods*. Cambridge, MA: The MIT Press.

Rogers, S. (2014). Insights into the #WorldCup conversation on Twitter. Retrieved from https://blog.twitter.com/2014/insights-into-the-worldcup-conversation-on-twitter

Schnapp, J., & Presner, T. (2009). The Digital Humanities Manifesto 2.0. Retrieved from http://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf

Somers, J. (2014). How I Reverse Engineered Google Docs to Play Back Any Document's Keystrokes. Retrieved from http://features.jsomers.net/how-i-reverse-engineered-google-docs/

Sudweeks, J., & Simoff, S. J. (1999). Complementary Explorative Data Analysis: The Reconciliation of Quantitative and Qualitative Principles. In S. Jones (Ed.), *Doing Internet Research: Critical Issues and Methods for Examining the Net* (pp. 29–56). Thousand Oaks, CA: SAGE Publications Ltd.

Thelwall, M. (2014). Sentiment Analysis and Time Series with Twitter. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.), *Twitter and Society* (pp. 83–98). New York: Peter Lang Publishing, Inc.

Weller, K., Bruns, A., Burgess, J., Mahrt, M., & Puschmann, C. (Eds.). (2014). *Twitter and Society*.

Wilken, R. (2014). Twitter and Geographical Location. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.), *Twitter and Society* (pp. 155–168). New York: Peter Lang Publishing, Inc.

Witmer, D. F., Colman, R. W., & Katzman, S. L. (1999). From Paper-and-Pencil to Screen-and-Keyboard: Toward a Methodology for Survey Research on the

Internet. In S. Jones (Ed.), *Doing Internet Research: Critical Issues and Methods for Examining the Net* (pp. 145–162). Thousand Oaks, CA: SAGE Publications Ltd.

Zimmer, M., & Proferes, N. (2014). Privacy on Twitter, Twitter on Privacy. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.), *Twitter and Society* (pp. 169–182). New York: Peter Lang Publishing, Inc.

## Appendix references

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why We Twitter: Understanding Microblogging. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* (pp. 56–65). http://doi.org/10.1145/1348549.1348556

Luck, E., & Mathews, S. (2010). What Advertisers Need to Know about the iYGeneration: An Australian Perspective. *Journal of Promotion Management*, *16*(1-2), 134–147. http://doi.org/10.1080/10496490903574559

Ofcom. (2013). *Communications Market Report 2013*.

Rios, M. (2014). Penalty kicks, as seen through Twitter data. Retrieved from https://blog.twitter.com/2014/penalty-kicks-as-seen-through-twitter-data

Thurman, N. (2007). The globalization of journalism online: A transatlantic study of news websites and their international readers. *Journalism*, *8*(3), 285–307. http://doi.org/10.1177/1464884907076463

# APPENDIX 1: COLLECTING AND QUERYING DATA

When you use Twitter's streaming API, tweets are pushed almost immediately after they are written and it is impossible to go back in time and request older tweets. Thus, it is important to have a continuous and stable connection with Twitter's servers, because when your server is down for an hour you miss that hour of data. Another thing to consider is the uneven distribution of tweets per unit of time. Attention on Twitter centers around provoking or captivating moments, especially in the case of football matches. On such moments, extra processing power is needed.

## Cloud server

For these reasons, I chose to delve into cloud servers as they have nearly no downtime and can be relatively easily scaled up (i.e. adding more processing power) whenever needed. The *Amazon Web Services*' (AWS) free usage tier offers a continually running micro server with 30 GB of storage. Although a micro server only offers a small amount of CPU resources, short bursts of additional activity are allowed. When registration to the AWS is completed, an Elastic Compute Cloud (EC2) server has to be launched and connected to an Elastic Block Storage (EBS) storage volume. When the server is running and accessible through Secure Shell (SSH) or the AWS Command Line Interface, the software needed for making a connection to Twitter's API can be installed. See the AWS documentation for more detailed information.

To set up a connection to the Twitter API I used the Phirehose library, which can save data in batches and regularly checks if the list of keywords has changed and reinitiates the connection when needed. Phirehose also fulfils Twitter's requirement of throttling reconnection attempts after a fail. To force my script to keep running forever, I set up a *cronjob*[15] that ran every 5 minutes and checked whether an instance of the script was already running. If that check returned *false*, the connection to the Twitter API was restarted. In other words: the connection was always restarted within 5 minutes after a server crash (provided that the server is working again).

The result of having a near continuous API connection during the whole tournament entailed a dataset of over 100 million tweets, a lot more than anything I had worked

[15] A cronjob executes a certain script or command every *x* minutes

with before. Although I reduced it to half its size, handling a dataset this large promised to be one of the toughest challenges of my research. I looked into three possible solutions, all of which offer unique advantages: a cloud computing tool, a non-relational database, and a relational database.

## Google Bigquery

When I discussed my dataset and research plans with some tech-savvy friends they pointed me to Google BigQuery. With BigQuery you can store your data on Google's servers, which allows very fast querying. Almost any query I ran on BigQuery took less than 3 seconds. The software also features a great interface and some nice out-of-the-box methods for URL parsing and date and time comparisons. Another great advantage is that you can share a BigQuery database with anyone with an Internet connection and a Google account.

BigQuery's main disadvantage is its pricing. While storage is very cheap, each query you execute on your dataset costs money. When searching or selecting the tweet text field, the cumulative size of **all** tweet texts is counted and billed per gigabyte. To search for a specific word in a dataset of 53 million tweets cost around 4 gigabytes to run, or around 25 cents at the time I tried it. A workaround would be to store data in multiple tables, but this would eliminate the advantage over a relational database, which can also be queried fast when using smaller tables.

Other problems of using BigQuery are that all your data has to be uploaded to Google's servers in strictly formatted files, and that it is not possible to edit data after you have inserted it into your database.

## MongoDB

After trying Google BigQuery I also gave MongoDB a try. As opposed to BigQuery, MongoDB can be used on your own computer, for free. It is also much more user-friendly: not only does it accept JSON as its input (the format that the Twitter API outputs), it also does not require a clear table structure at all and new columns can be added at any moment. The disadvantage of MongoDB over BigQuery is shareability: to share a MongoDB database a web server is needed.

While MongoDB is supposed to be a lot faster than a relational database when querying big datasets I was not able to achieve this myself. When the size of your dataset (including indexes) exceeds the amount of memory your machine has, speed decreases drastically. Besides that, querying is not as easy as querying a relational database.

## MySQL

Finally, I tried the relational database management system MySQL. I had used MySQL for earlier research on smaller datasets. As it is still the most popular option for web developers, the advantages of MySQL include its wide availability and ease-of-use. Querying tables is very intuitive and data can be easily edited. I found that simple queries like searching for tweets by their ID were executed very fast.

Like with MongoDB, a disadvantage of MySQL is its shareability as you need to set up a web server. The biggest caveat of MySQL, however, is the speed of textual searching. Searching the complete set of tweets for a certain word or phrase takes minutes. Because of this, I decided to combine MySQL with a non-relational database system made specifically for textual searching: ElasticSearch. Searching for a specific keyword takes a few seconds only, even when results are filtered and/or aggregated (useful when creating a timeline of results). For every tweet I added its ID, timestamp and text to an ElasticSearch database. A search returned all matching IDs, which I then looked up in the MySQL database.

## Conclusion

Choosing which database management system to use depends on the data and the research group. When working on the same dataset with multiple people I would advise to use software like BigQuery, as it can be made available to anyone with a Google account. When needing to do research on a very short notice or even in real-time I would suggest using MongoDB because of the ease of inserting new records. In other cases, and especially when there already exists some knowledge of SQL, I would advice MySQL, possibly in combination with ElasticSearch.

# APPENDIX 2: TIMELINES

One of the first things I always do after I have collected a sample of tweets is to create a timeline of the number of tweets per minute (or any other unit of time). I simply round the timestamp of each tweet to the nearest minute and then calculate the sum of tweets for each minute. With very popular subjects, however, tweets may be omitted due to API limitations. This renders a timeline invalid, since high peaks are flattened out to the maximum number of tweets the API allowed. Thankfully, Twitter provides information about the number of tweets that were omitted. Whenever an omission occurs, they post a *rate limit message* containing the cumulative number of tweets that were left out until that moment.

I collected all these messages and took the maximum value for each minute. This gave me a list of timestamps and the cumulative number of omitted tweets up until (not *in*) that minute. After subtracting the previous value for each minute, I got the absolute number of tweets omitted for a given time. Since you then know the total number of tweets at any given time and what percentage of these tweets is actually contained within the sample, each timestamp can be given a multiplier:

| Timestamp | In sample | Omitted | Multiplier | About subject | Estimation |
|---|---|---|---|---|---|
| 3-4 | 200 | 20 | 1.1 | 10 | 11 |
| 4-5 | 200 | 100 | 1.5 | 8 | 12 |

**Table 2 – Rate limit timeline correction**

The last two columns of table 2 represent the actual number of tweets about a subject that are in the sample, and an estimation of the total number of tweets about a subject, based on rate limit messages. This is not only useful when creating a timeline, but also when comparing searches between timestamps.

Please note that although the above presents a good method for making comparisons within the data sample, it says nothing about the sum of all tweets that were written about the subject. There are always tweets that do not contain any of the tracked keywords; for example ‘opening ceremony’ or any player’s name. I compared the numbers of my own data sample[16] to a few figures Twitter published on their blog[17] and

[16] I missed some data for matches Bra vs. Chi and Col vs. Uru, so these cannot be compared

found out they do not match consistently; the percentage of tweets I obtained is much higher for some matches than for others:

| **Match** | **Twitter** | **My data** | **Percentage** |
|---|---|---|---|
| BRA vs. GER | 35,6 million | 3,6 million | 10,1 |
| NED vs. ARG | 14,2 | 2,6 | 18,3 |
| BRA vs. COL | 12,4 | 2,0 | 16,1 |
| ESP vs. NED | 8,3 | 1,1 | 13,3 |
| GHA vs. USA | 4,9 | 1,1 | 22,4 |

**Table 2 – Number of tweets per match, in total and in my sample**

I have no explanation for this, except that people from certain countries may use hashtags or hashflags more often than others. Also, people may have used unexpected variations of the 'official' match hashtag: in the case of Spain vs. Netherlands, a lot of Dutch people used the hashtag *#SPANED* instead of *#ESPNED*. The nature of the match may also be of influence, with highly emotional matches making people 'forget' about using a hashtag.

[17] https://blog.twitter.com/2014/the-twitter-worldcup-group-stage-recap, https://blog.twitter.com/2014/insights-into-the-worldcup-conversation-on-twitter, https://blog.twitter.com/2014/seven-worldcup-data-takeaways-so-far

# APPENDIX 3: MEDIA CONVERGENCE

During Spain vs. Netherlands, a match that I, being Dutch myself, remember vividly, ESPN posted five tweets. Four of them contained photos: one was a graphic announcement of the kick-off/broadcast time and the other three were screenshots of the ESPN broadcast. One shows the moment the ball passes the goalkeeper after the penalty kick, showing how close it really was, and another one is of Arjen Robben posing for a camera after he scored a goal. The only link that ESPN shared was directed to their own website (espn.com), to a slow-motion replay of the moment Spain was rewarded a questionable penalty kick (unfortunately the video is deleted now).

Interestingly, these photos and link seem to *add* something to ESPNs television broadcast. It indicates that ESPN utilizes Twitter as an additional platform to fulfil the needs of 21st century media consumers, offering a second screen where people can get supplementary footage or information. In this case study, I will take a closer look at tweets shared by accounts of media organisations and try to answer whether they are using Twitter as an addition to their other activities, indicating a sign of media convergence and a "flow of content across multiple media platforms" (Jenkins 2006)?

## Methodology

I first identified accounts of media organisations by filtering out accounts that meet the following requirements:

- The profile bio contains a hyperlink
- The account has at least 100.000 followers
- The account is verified
- At least 50 statuses about World Cup matches

This resulted in a set of accounts of media, journalists, celebrities, sports clubs, and a few other people and companies. I then determined which of these accounts could be regarded as 'media' using the account descriptions, which often contained words like 'news', 'coverage' or 'sports'. I chose to exclude Arabic and Asian accounts because of their readability, and ended up with a selection of 115 broadcasters, newspapers and online media.

After that, I made a table containing all tweets by these accounts. This resulted in a dataset of 53.814 tweets, 55% of which contained an URL[18]. Using a PHP script I resolved shortened URLs (using services like ow.ly and bit.ly) to their full counterparts. After the removal of a few URLs due to a failed address lookup (for example because of a typo in the short URL), this resulted in a set of 29.823 shared URLs[19].

## Results

Comparing the domain name of those URLs to the ones as added to the account profiles resulted in 21.443 matching domain names, meaning around 72% of shared links pointed to the medium's own website. However, I knew that of the other links, the majority was between different websites/accounts of the same company. Organisations like ESPN and BBC use a variety of accounts and websites, as they cater to multiple countries and languages (BBCWorld, ESPNDeportes) or have accounts dedicated to sports (BBCSports) or TV shows (SportsCenter). To tackle this problem, I determined the company name of each of the shared domain names.

Using this method, only 1401 (4,7%) of the URLs linked to a website not owned by the tweet author's organisation. Most of these links directed to social media websites like Vine, Facebook, Twitter, Youtube or Instagram. The media organisations mostly linked to their own accounts on those websites, and on some occasions to posts made by celebrities or athletes. After I excluded those websites, only 123 links were left. Of these, 52 linked to the official FIFA website, and 42 directed to various online betting websites (all posted by Mirror.co.uk). So finally, I found only 29 links, or less than 0.1%, that did not point to one of the author's own websites.

It appeared some media organisations set up their Twitter account to automatically republish their own liveblogs/headlines, like @NOSvoetbal. This allows followers to "get in-the-moment updates" and "watch events unfold in real time", matching the purpose of Twitter as described on the platform's own home page. It also serves as a gateway for attaining more information, which benefits both the consumer and the producer.

[18] The tweets that did not contain a link seem to be live commentary in most cases, containing information about starting squads, goals, substitutes, chances or match statistics.

[19] All shared links can be accessed through http://thesis.martijnweghorst.com/media/links

***Retweets***

Of the 53.814 media tweets, 14.583 (~27%) proved to be a retweet. Following the previous conclusion about the lack of links to different websites, I expected these retweets to be cases of media companies having multiple accounts and retweeting themselves, like BBCSport, BBC, BBCWorld and bbcmundo, or ESPN, ESPNDeportes and SportsCenter. Unfortunately, I did not come up with a way to automatically identify retweets that existed outside of a company's own Twitter-sphere. What I did instead was counting he number of different users each media account retweeted.[20]

The results show that most media barely post retweets, as they retweeted 10 different users on average and 80 accounts only retweeted 5 users or less. Also, these retweets appeared to be of different accounts owned by the same medium or of journalists employed by this medium (for example Welt). There were only a few exceptions of accounts that regularly retweeted 'regular' users (for example Sky, ZDF, and Zeit).

## Conclusion

In conclusion, I have made clear that traditional media attempt to use Twitter as an addition to their original activities. More than half of the tweets sent out by the scrutinized media organisations contained a link, and with little exception these links directed to the company's own website. Furthermore, their retweets often link to other accounts owned by them. In further research, it would be interesting to do a social network analysis and study whether there exist distinguishable small Twitter-spheres of connected accounts.

There also seems to be a big difference between the accounts of media *companies* and *journalists*, as the latter share relatively few links. I created a set of tweets by 15 individual journalists in order to study whether they only linked to the website of their employer. But whereas media accounts shared links in 55% of their tweets, journalists did so in only 3%. More research could study the Twitter activity and intentions of those individual journalists.

[20] The complete table and all retweets can be viewed on http://thesis.martijnweghorst.com/media/retweets

# APPENDIX 4: MEDIA GLOBALISATION

In this case study, I will look at media consumers rather than producers. Thurman (2007) found that 36% of the visitors of British news websites come from the United States, and an additional 39% from other countries outside of the UK. I will do a similar study using big data and try to find out to what extent media are read by an international audience.

My dataset of World Cup tweets is particularly suitable for this purpose as we can assume that people all over the world are interested in its subject, and that most national news media have coverage of it.[21] I think that especially Spanish and English media will be interesting to look at. There are a lot of Spanish speaking countries and even though citizens living in these countries speak the same language and could in theory all simply follow the Spanish website of ESPN, they each have their own national media. English is the most popular second language, which could greatly benefit American and British media like CNN and BBC.

## Methodology

To answer questions about globalisation, I needed geographic information of Twitter users. Unfortunately, the API information that indicates locations is often ambiguous:

- The tweet and user language are not bound to a country. Furthermore, the tweet language is detected automatically and the user language is simply a user preference;
- The time zone selected by a lot of users appears to be the alphabetically first time zone that corresponds to their own in terms of difference to GMT (Swiss users selecting Amsterdam instead of Zurich);
- The most reliable are a tweet's coordinates, but a user has to explicitly allow Twitter so save it. In this dataset, only 3,65% of tweets (1.935.461) contained coordinates. Also, retweets posted by clicking on the retweet button never have coordinates attached to them.

[21] For this reason, other datasets may give other results. When studying a subject like national politics or Major League Baseball, results could be different either because only Americans are interested, or because only American media cover it (sufficiently).

Because I needed a bigger sample of geo-tagged tweets and, more importantly, one containing retweets, I chose to automatically identify the geographic location of tweets by parsing the user location text and match on country names. I used a list with country names in both English and in the language(s) spoken in that country, (for example Schweiz, Suisse and Svizzera as alternatives for Switzerland). I excluded the Maldives and Sri Lanka because of difficulties with their script, and the Congo's because of possible ambiguity. For the full list of country names, see Appendix 7. After parsing, I removed tweets where the user's location contained more than one country name.

The scripted detection resulted in a corpus of 11.552.082 tweets (of which 45% were retweets) for which I had identified the location of its author, nearly six times as big as the number of tweets with coordinates. I checked the reliability of my method by comparing these tweets to the ones with explicit coordinates and found that 91,1% matched. One of the causes of non-matches is that location texts can be ambiguous: Armenia, for example, is not only the name of a country but also of a city in Colombia. Other causes are that people are only visiting Brazil temporarily, and that organisational accounts dedicated to the World Cup locate themselves in Brazil regardless of the real location of their writer/owner. This reveals an unanticipated advantage of my method: it measures nationality, whereas Twitter returns a user's current location.

Besides looking at the accounts that were most retweeted per country, I also looked at shared links. Contrary to the previous chapter I decided not to unshorten URLs, as this sample contained over 200.000 unique URLs. At first I identified the top-level-domain (.co.uk or .com) for each shared link, so I could count the number of .co.uk links in the Netherlands (for example). However, as there are a lot of URL shorteners like cnn.it, and because websites often use the .com domain regardless of their location (voetbal.com, for example), this method lacked reliability. So instead, I simply identified the top *hosts* (wsj.com, bbc.in) per country.

One of the issues of identifying countries from a user's location text is that its success rate is not evenly distributed over countries. American citizens, for example, set their user location to a combination of city name and state abbreviation. This causes comparisons of absolute numbers between countries to be invalid and renders claims about the geographic distribution of particular accounts or websites useless. Instead, *rankings* of the most retweeted accounts and shared domains can be made, per country and based on unique users. Consequently, I could see the ranking in each country per account and domain, giving an indication of their global popularity.

***Bit.ly statistics***

In addition to shared links and retweets, I also looked at bit.ly click statistics. Bit.ly is an URL-shortening service that is used often when text space is limited (consider Twitter's 140-character limit). When someone clicks on a bit.ly link, he is redirected through the bit.ly server to the intended URL. During this redirection, bit.ly saves metadata about the user, like his geographical location. Bit.ly makes these statistics publicly available through its API and website (simply append '+' to a bit.ly URL).

Some of the domains that were shared most often, like es.pn and bbc.in, appeared to be aliases of bit.ly. To obtain a list of some of these aliases, I grouped all shared links based on their domain name and counted the occurrences and the number of unique users that shared links to this domain. I then checked for each domain whether or not it was a bit.ly alias. Media that used a bit.ly alias included ESPN, BBC, L'Equipe, Washington Journal, Globo and some other (mostly South American) media.

I copied all the links to one of these bit.ly aliases to a new table, which resulted in a dataset of 8.299 unique URLs. I then proceeded to download the number of clicks per country for each of these URLs. Because some bit.ly aliases point to different websites (often because media organisations offer their content in multiple languages, like fr.fifa.com or huffingtonpost.es), I decided to base my study on full domain names rather than the bit.ly pro domain. I chose to include all domain names that occurred more than 30 times, resulting in a set of 50 websites. For each of these websites, I filled in the country of its owner and calculated the percentage of viewers from that same country ('domestic clicks'). So for ESPN Deportes I calculated the percentage of visitors that came from within the US (the website is Spanish and caters to Hispanics in the US), for the BBC from the UK, etcetera. Finally, these results present a given media organisation's percentages of domestic and foreign readers.

## Results

When looking at the accounts that were most retweeted in a country, the US and UK show similar patterns. The top 3 of both countries consist of the official FIFA account, the national team's account, and one media account (SportsCenter and BBCSport, respectively). The most popular American medium in the UK is ESPNFC on place 11, and the most popular British medium in the US is BBCSport on place 13. The BBC appears to be relatively global, being that they appear in the top 25 of retweeted accounts of 80 countries. Furthermore, BBC.in is ranked in the top 25 of all 111 studied countries most shared links lists, except for Mexico (26th). For all rankings of @BBCSport and bbc.in, see http://thesis.martijnweghorst.com/accounts/BBCSport and http://thesis.martijnweghorst.com /websites/bbc.in.

What is interesting, and contrary to what I expected, is that there are no Spanish-language media that are ranked high in a lot of South American countries' retweet rankings. Medio Tiempo, the top Mexican medium, appears only in one other top 10, in Bolivia. A lot of South American countries appear to have one or two **national** media organisations whose tweets are retweeted by most users – for example Costa Rica (Nacion), Ecuador (marcadorec), Mexico (Medio Tiempo, Sopitas), Colombia (El Tiempo), and Venezuela (UNdeportes). What this tells us is that people prefer to follow a national medium, at least in the case of football. Of course this has something to do with coverage of the national team; we can assume Ecuadorian media have better or more coverage of their national team than Mexican and Chilean media do, despite their common language.

### *Domestic clicks*

The results of the bit.ly click analysis are as I expected: the websites with the smallest percentages of domestic clicks are American-based websites catering to the Spanish community. The rest of the top 10 consists of language independent websites like Wall Street Journal Graphics and Huffington Post Data[22] and American and British websites.

| | | Website | Links | Clicks | Domestic |
|---|---|---|---|---|---|
| 1 | 🇺🇸 | cnnespanol.cnn.com | 43 | 1207204 | 17% |
| 2 | 🇺🇸 | espndeportes.espn.go.com | 186 | 137862 | 24% |
| 3 | 🇺🇸 | foxdeportes.com | 171 | 1639585 | 30% |
| 4 | 🇺🇸 | voces.huffingtonpost.com | 69 | 1063 | 35% |
| 5 | 🇺🇸 | edition.cnn.com | 47 | 198741 | 36% |
| 6 | 🇬🇧 | dailymail.co.uk | 200 | 155903 | 40% |
| 7 | 🇺🇸 | graphics.wsj.com | 41 | 21953 | 40% |
| 8 | 🇺🇸 | online.wsj.com | 49 | 172438 | 47% |
| 9 | 🇺🇸 | data.huffingtonpost.com | 45 | 35491 | 48% |
| 10 | 🇬🇧 | bbc.co.uk | 961 | 905539 | 49% |
| 11 | 🇫🇷 | fr.sports.yahoo.com | 118 | 14244 | 51% |
| 12 | 🇬🇧 | huffingtonpost.co.uk | 96 | 53885 | 57% |

[22] For examples, see http://graphics.wsj.com/documents/WORLDCUPTOEE/, http://graphics.wsj.com/wc-game-recaps/?mod=e2tw#/?g=731817, and http://data.huffingtonpost.com/2014/world-cup/matches/brazil-vs-netherlands-731829

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| 13 | 🇺🇸 | cnn.com | 55 | 82419 | 63% |
| 14 | 🇻🇪 | brasil2014.ultimasnoticias.com.ve | 36 | 19081 | 64% |
| 15 | 🇺🇸 | msn.foxsports.com | 101 | 36095 | 66% |
| 16 | 🇻🇪 | ultimasnoticias.com.ve | 184 | 309043 | 67% |
| 17 | 🇺🇸 | blogs.wsj.com | 146 | 188470 | 67% |
| 18 | 🇺🇸 | sports.yahoo.com | 144 | 91561 | 67% |
| 19 | 🇦🇷 | infobae.com | 214 | 157985 | 72% |
| 20 | 🇺🇸 | espnfc.com | 189 | 930438 | 72% |
| 21 | 🇻🇪 | liderendeportes.com | 258 | 126325 | 73% |
| 22 | 🇻🇪 | brasil2014.liderendeportes.com | 75 | 79916 | 73% |
| 23 | 🇺🇸 | huffingtonpost.com | 116 | 385617 | 74% |
| 24 | 🇪🇸 | huffingtonpost.es | 33 | 5267 | 74% |
| 25 | 🇲🇽 | record.com.mx | 189 | 149293 | 77% |
| 26 | 🇲🇽 | mediotiempo.com | 241 | 1001232 | 78% |
| 27 | 🇧🇷 | oglobo.globo.com | 219 | 196830 | 78% |
| 28 | 🇲🇽 | mmdeportes.telediario.mx | 94 | 6478 | 80% |
| 29 | 🇺🇸 | espn.go.com | 400 | 504504 | 82% |
| 30 | 🇺🇸 | worldsoccershop.com | 46 | 30460 | 82% |
| 31 | 🇲🇽 | laaficion.milenio.com | 172 | 215780 | 83% |
| 32 | 🇫🇷 | lequipe.fr | 121 | 1832348 | 83% |
| 33 | 🇫🇷 | huffingtonpost.fr | 46 | 12662 | 85% |
| 34 | 🇪🇸 | telecinco.es | 100 | 1309982 | 88% |
| 35 | 🇲🇽 | sopitas.com | 46 | 803247 | 89% |
| 36 | 🇧🇷 | epoca.globo.com | 32 | 3776 | 89% |
| 37 | 🇧🇷 | esportes.estadao.com.br | 114 | 1406987 | 90% |
| 38 | 🇧🇷 | g1.globo.com | 39 | 202200 | 92% |
| 39 | 🇧🇷 | cbn.globoradio.globo.com | 33 | 21956 | 93% |
| 40 | 🇧🇷 | globoesporte.globo.com | 280 | 2700465 | 94% |

Spanish-speaking countries have a lower percentage of domestic clicks than French and Brazilian websites do. This can be explained as follows: the more foreigners speak a language, the bigger the international market for a website in that language is, and French and Brazilian are spoken in fewer places than English and Spanish. That is also why Venezuelan media have a higher percentage of foreign readers than Mexican media; Mexico has a population four times as big. The same logic can be applied to the difference between English and Spanish media; a lot of people speak English as a second or third language.

The Wall Street Journal (WSJ) scores a very low domestic click percentage. This is mostly caused by clicks from The Netherlands (almost 28%). These Dutch clicks entailed two WSJ articles in particular, both discussing the success of the Dutch squad. The first one was written after the 5-1 win over Spain and addresses the impact of the popularity of field-hockey in The Netherlands on its football tactics. The second article was written after the US team was eliminated and is an open letter with reasons to support the Dutch squad for the rest of the tournament. Of course, the fact that precisely these two WSJ articles were so popular among Dutch people shows that people are willing to link to foreign media when the content is relevant to them; in this case giving them a feeling of pride. Had the article promoted the German squad, it likely would have been linked to mostly from Germany rather than from Holland.

What is interesting about the South American Spanish websites is that for each the country with the second most clicks is the USA (by a distance), and that Spain always ranks 3rd or 4th, above a lot of South American countries. This may be because media from the US and Spain don't have enough coverage of South American teams, causing expats and other citizens interested in information pertaining to these teams to look elsewhere.

## Conclusion

In this case study I have presented two methods for studying media globalisation. Firstly, I analysed which media organisations are most retweeted in a given country. Because of an uneven distribution of geo-tagged tweets, this analysis did not disclose where the majority of followers of a certain media organisation came from. For this reason, I also studied bit.ly data to try and find out whether certain websites attract more foreign readers than others.

Both the retweet and bit.ly data analysis confirmed that whenever people engage with foreign media, they still often do so in their native language. ESPN is popular in Canada, Australia and India, whereas l'Equipe is visited often from Algeria and

Morocco. The bit.ly analysis also showed that English media are more likely to attract non-domestic readers compared to Spanish media, and that CNN is the most globally popular medium. This could be a result of their global broadcasting, since people are already familiar with CNN because of their offline presence. The example of the WSJ articles that were clicked on in The Netherlands showed that whenever a lot of people do read foreign media content in a different language than their own, the content is extra relevant to their interests.

## Discussion

I have to admit that it was difficult to draw conclusions from this data. The method using retweets was influenced by inconsistent hashtag use: since I tracked only certain hashtags when constructing my dataset, only accounts correctly using hashtags are present in the rankings. A lot of the most popular tweets may not have contained a hashtag and therefore have not been captured within my dataset. It is conceivable that there is a more global media organisation but that it simply did not use hashtags. Reversely, it may be one of the reasons the @FIFAWorldCup is so high in all popularity rankings, as they **did** properly use hashtags in most of their tweets.

For future research, it might be interesting to use a dataset about a subject that is not covered globally, for example local politics, and study whether Dutch people read more American media when the subject is American politics.

# APPENDIX 5: MEDIA MESHING

Whereas I studied the way media organisations use Twitter as an addition to other activities in Appendix 3, I will now look into the way media *consumers* use Twitter as an addition. Media consumers can use a combination of devices and technologies simultaneously in order to enrich the media experience, also known as media meshing:

> *The iYGeneration finds itself faced with more tasks on a daily basis […] and have become highly proficient at multi-tasking and media meshing (i.e. consuming one or more media at once). Media meshing is a behavioral trend, exemplified by simultaneously watching television, surfing the Internet, listening to iTunes music, and texting, while traditional media is often pushed to the "background". This behavior is explained as a constant search for complementary information, different perspectives, and even emotional fulfillment*
> Luck and Matthews (2010)

In Ofcom's annual Communications Market Report (2013), it is reported that a quarter of UK adults regularly mesh media, with texting and making phone calls about television programs being the most common activities.

Java et al. (2007) claim that "people use microblogging to talk about their daily activities and to seek or share information", making Twitter an excellent platform for media meshing. The dataset of tweets written during World Cup matches may be particularly interesting, as we can assume a lot of these Twitter users were simultaneously watching a television broadcast or listening to the radio. This is more or less confirmed by Rios (2014), who shared a timeline of Twitter use during penalty shootouts. What his analysis shows is that people are very silent when the player is getting ready, and then massively respond when the penalty kick is converted or missed. This is reminiscent of the offline situation: people tend to be quiet and concentrated during the build-up and then start displaying their relief, happiness or disappointment. The graph practically confirms media meshing, since it is virtually impossible to be so up-to-date with real-time events without using another, live, medium.

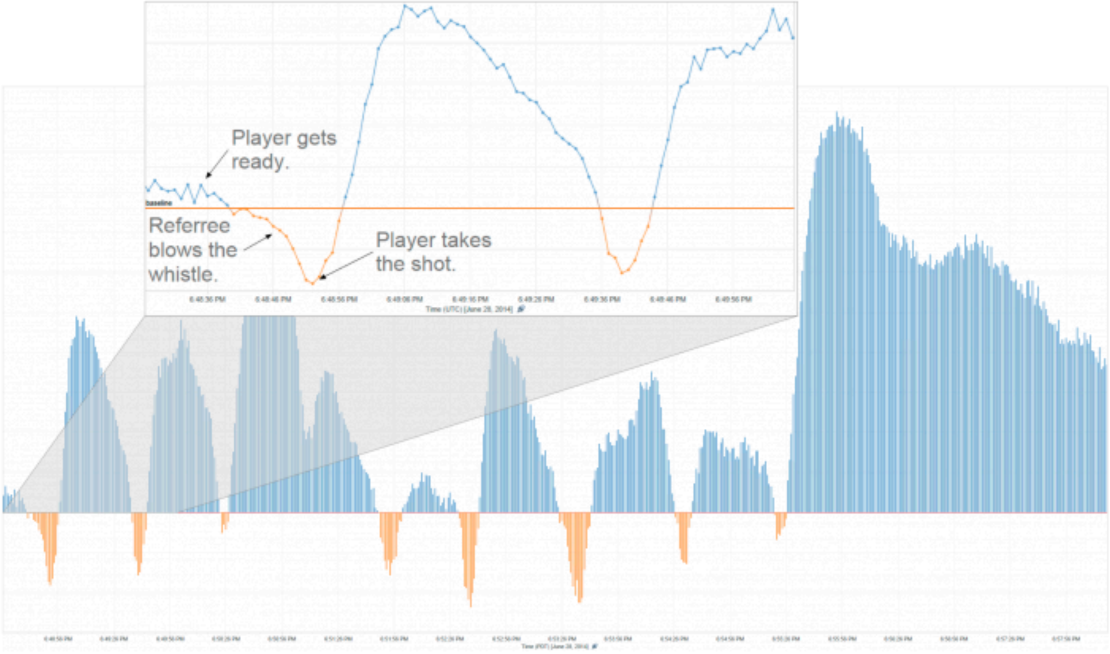


**Figure 2 – Twitter use per second during penalty shootouts (Rios 2014)**

In the remainder of this appendix I will study how match progress affects Twitter use. Are there differences in percentages of retweets, replies, shared images or mobile use? Are people reporting results ('news' that is probably known already by everybody who is interested in it), or are they merely expressing emotion for personal satisfaction? Finally, I'm hoping to give some insight into the specifics of Twitter use during major sports events.

## Methodology

Firstly I needed to link tweets to the corresponding moment in the match, using the timestamps of all tweets. As opposed to Rios, I did this on a per-minute basis. Twitter's rate limiting resulted in the fact that for a lot of seconds I retrieved the exact amount of 200 tweets. Although the streaming API gives information about the number of tweets that was omitted, it seemed it did not do so on a correct per-second basis.

To be able to link tweets to events, I needed factual data about the matches, such as when goals were scored of red cards were given. I obtained this data by downloading all 64 live blogs from FIFA.com. I then saved this data to a table where each row consists of a timestamp, a match number and the type of event. As the FIFA liveblogs contained information about kick-off, breaks, and the final whistle, I was also able to create a list with all the minutes of each match (actually from one hour before kick-off until one hour after the final whistle) and whether the respective minute was pre- or post-game, during a break, or during play (subdivided in first half, second half, overtime, and penalty shootout).

For all these minutes and match parts I calculated how many tweets were written. As was to be expected, the number of tweets per minute was the highest right after goals were scored and when the referee blew the final whistle. Other peaks were caused by disallowed goals[23], awarded penalties (regardless of the outcome[24]), and red cards[25].

Because studying the plain number of tweets per minute in and of itself does not grant insight into the motivations for using Twitter, I identified various tweet characteristics. I distinguished retweets and conversations by looking at whether the tweet text started with "RT @“ or “@”, respectively. I categorised tweets in mobile vs. non-mobile and in iOS and Android groups by looking at all distinct values of the ‘source’ field and manually determining their hardware and software. I separated tweets with links, images and mentions by looking at the ‘entities’ fields. I excluded retweets from the mentions.

Using all this data, I created graphs showing the progress of those percentages per match, with important match events annotated at the bottom (see Appendix 8 for an example). In order to be able to evaluate whether phenomena not only occurred during single matches, I also calculated averages per match part, showing patterns like the percentage of retweets and mobile/iOS/Android use, mentions, replies, images and links. These are averages of the priorly calculated **percentages**, so as not to skew the average towards knockout-stage effects (because more tweets were written during that stage of the tournament).

## Results

Table 4 shows that the percentages of tweets that are retweets or contain a link are lower when the match is not in progress. This indicates that during matches, people use Twitter in a relatively egocentric fashion and as a way to broadcast their own emotions or opinions rather than for following other people. This is confirmed by the fact that these percentages decrease even more right after goals. During half time, when television channels show advertisements, people have time to read and share some tweets or links.

[23] Three in the first half of Mexico vs. Cameroon

[24] France vs. Switzerland 30th minute

[25] Germany vs. Portugal 35th minute

|  | Pre | 1st half | Pause | 2nd half | Overtime | Penalties | Post |
|---|---|---|---|---|---|---|---|
| **Android** | 28.9% | 32.1% | 31.8% | 31.2% | 35.1% | 33.8% | 33.3% |
| **Replies** | 2.1% | 1.7% | 2.0% | 1.5% | 1.8% | 0.83% | 1.7% |
| **Avg. length** | 95 | 81 | 95 | 83 | 81 | 57 | 102 |
| **Foursquare** | 0.032% | 0.020% | 0.011% | 0.006% | 0.004% | 0.002% | 0.002% |
| **Images** | 32.3% | 21.0% | 33.8% | 18.7% | 17.5% | 4.8% | 35.6% |
| **Fans** | 6.1% | 5.0% | 6.8% | 5.2% | 4.8% | 2.2% | 7.3% |
| **Instagram** | 0.40% | 0.25% | 0.30% | 0.17% | 0.14% | 0.037% | 0.29% |
| **iOS** | 33.4% | 34.4% | 33.3% | 33.6% | 37.2% | 40.0% | 33.6% |
| **Links** | 9.1% | 11.0% | 15.6% | 11.6% | 9.0% | 9.5% | 15.6% |
| **Mentions** | 17.2% | 13.0% | 17.2% | 12.7% | 11.0% | 5.6% | 19.5% |
| **Mobile** | 72.7% | 74.4% | 75.9% | 73.3% | 77.9% | 74.4% | 75.4% |
| **Negative** | 0.35% | 0.66% | 0.66% | 0.92% | 1.2% | 1.1% | 1.0% |
| **Positive** | 1.5% | 1.4% | 1.2% | 1.3% | 1.2% | 0.88% | 1.6% |
| **Retweets** | 54.1% | 46.9% | 63.8% | 46.4% | 47.3% | 33.9% | 67.4% |

**Table 4 – Tweet characteristics over the course of a match**

The average length of tweets also decreases when the match is in progress (and especially during penalty shootouts). This confirms Twitter's premise of getting in-the-moment updates and watching events unfold in real-time as people are sharing match events as quickly as possible:

> *Compared to regular blogging, microblogging fulfills a need for an even faster mode of communication. By encouraging shorter posts, it lowers users' requirement of time and thought investment for content generation.*
> Java et al. (2007)

For some variables, there was a distinction between the group stage and the knockout phase. For example, tweets written during the knockout phase were 18% more likely to be a direct message, meaning people are having more conversations. The number of tweets for which I identified a sentiment was also higher during the knockout phase, suggesting a higher level of emotional involvement.

### *Fans versus neutral viewers*

Using the method I deployed for analysing globalisation of media readers, I also compared fans and neutral viewers. When a user's location text contained a country name, this user was labelled as a fan of that country.[26] The seventh row ('Fans') in table 4 shows that the fluctuation of the percentage of tweets that is written by fans follows a pattern: when the match is not in progress, that percentage is a lot higher. Netherlands versus Argentina provides a clear example of this:

|  | Pre | 1st half | Pause | 2nd half | Overtime | Penalties | Post |
|---|---|---|---|---|---|---|---|
| **NED vs. ARG** | 6.7% | 4.5% | 6.6% | 4.1% | 4.8% | 2.3% | 8.2% |

**Table 5 – Percentage of tweets by fans during Netherlands versus Argentina**

During each moment wherein no play is taking place the percentage of fan-engagement is higher. This indicates that fans are paying more attention to their television compared to non-fans on moments when the match is being played. However, a closer look revealed that although the *percentage* of fans is lowest during the penalty shootouts, the *absolute* number is actually higher than during the majority of the match. This means that the increase during these exciting minutes occurs in both groups of watchers, but relatively more so for non-fans. This is a logical phenomenon: people who are less involved have a higher threshold of posting something. This means that relatively many neutral viewers wait to join in on the discussion until something important happens, whereas fans were already twittering anyway. Looking at the moments of goals confirms this theory: after both goals scored during Brazil versus Chile there was a slight drop in the *percentage* of fan tweets, whereas the *absolute* number of fan tweets increased. This also explains pre- and post-game percentages: the build-up and aftermath is simply more interesting when you are emotionally involved.

### *Mobile*

I found that almost all of the matches in the knockout stage had a relatively high percentage of mobile use. This led me to calculate the percentages of mobile use in both the group and knockout stages: 74,8% and 77,6%, respectively. This *could* have something to do with people watching in bars/groups more often when the tournament progresses, thus grounding a sociological claim like that in web data. However, there

[26] Unfortunately I could not compare absolute numbers between matches because of this, as explained earlier. Matches of the USA would have a disproportionately low number of fans, as US inhabitants are not used to putting USA in their location text.

could also be other explanations, like a difference in mobile use between countries that progressed and were eliminated.

What would have been interesting is to measure mobile use over the course of a whole day, in order to find out whether mobile use increases towards the beginning of a match. This could potentially have other reasons as well, though – for example the fact that Europeans generally would be at work (behind a desktop computer) during the day and at home (on the couch with a tablet) at night.

### *Instagram links*

One of the things I noticed when analysing shared links for the previous chapter was that Instagram was the website that most Dutch and German users linked to. As Instagram is a platform where people mostly share photos of themselves and/or made by themselves, this made me wonder what the incentive was to do this during (or right before/after) a World Cup match, and whether images posted to Instagram are different from images shared on Twitter.

| | Pre | 1st half | Pause | 2nd half | Overtime | Penalties | Post |
|---|---|---|---|---|---|---|---|
| **Images** | 32.3% | 21.0% | 33.8% | 18.7% | 17.5% | 4.8% | 35.6% |
| **Instagram** | 0.40% | 0.25% | 0.30% | 0.17% | 0.14% | 0.037% | 0.29% |

**Table 6 – Percentage of tweets with images and Instagram links**

Although the differences between images shared directly on Twitter and links to Instagram.com are small (relatively speaking, of course the dataset contains much more Twitter images), there is one thing that caught my attention. Twitter images are shared approximately the same number of times during pre-game, half time, and post-game discussion. Instagram links, on the other hand, are shared 30% more often during pre-game talk than during half time or after the game.

I did an assessment of 40 post-match images on both platforms (using England versus Italy and Brazil versus Germany) and this delivered the following results: of the 40 Instagram links, 24 directed to match photos or collages or memes and 16 linked to personal photos. Of the 40 Twitter photos none were a selfie and only two were personal photos: one of a television and the other one on the streets. All of the other links directed to either screenshots, Photoshopped images, match results graphics, or match photos. What this closer look at individual data points thus revealed is that the types of images that are shared on Instagram and Twitter are very different.

## Conclusion

In this final case study, I again attempted to mobilize a Twitter dataset in order to answer a prevailing question within media studies. I studied why and how people simultaneously watch a television broadcast and use Twitter, also known as media meshing. In order to be able to do so, I linked tweets to FIFA.com liveblogs, the most primary source available for match coverage. In order to show how Twitter use changed during a match, I created a table with certain tweet characteristics plotted against match parts (pre-game, first half, break, etcetera). The result is a table that contains the averages of all 64 matches.

Looking at the percentages of retweets and links revealed that reasons for media meshing can change over the course of a match. When a match is in progress, Twitter serves as an emotional outlet or as a way of broadcasting news, while during pauses it is a medium for getting additional information.

Drawing more specific conclusions from the table turned out harder than it first seemed. For example, the difference between absolute versus relative numbers can be misleading. Upon first look, one could take away that fans of the participating teams are using Twitter mostly when the game is not in progress, as the 'Fans' figure drops when the ball is in play. However, the table displays the *percentage* of tweets that is written by fans. In reality, both fans and neutral viewers tweet more during the match. The alternative and more likely explanation is that neutral viewers have a higher threshold of tweeting about a match: they only do so when something important happens.

In order to really draw conclusions, looking deeper into the dataset is necessary. The table showed that Instagram links were shared the most during pre-game talk, whereas Twitter images were shared equally often during pre-game, pause, and post-game talk. A closer look, for which I studied 80 images, revealed that Instagram is a platform where people share relatively many photos of themselves and/or made by themselves, whereas on Twitter people mostly share memes and screenshots.

# APPENDIX 6: DEVICES

This is a list of the 100 most-occurring ‘sources’ as returned by the Twitter API. The two rightmost columns display whether the device is mobile, and on what platform it runs.

Not all of these sources can be clearly categorised. Hootsuite (#14), for example, is available as a website, as an iOS app, and as an Android app. Furthermore, personal knowledge is required as not all these sources are self-explanatory. I only found that Tweetlogix (#51) is actually an iOS app after I had done my research. It would be useful to collectively create a list of known ‘Twitter sources’ and their characteristics.

| | Name | Tweets | Mobile | Platform |
|---|---|---|---|---|
| **1** | Twitter for iPhone | 16.777.215 | ✓ | iOS |
| **2** | Twitter for Android | 13.364.932 | ✓ | Android |
| **3** | Twitter Web Client | 9.067.680 | | |
| **4** | Twitter for  Android | 2.354.783 | ✓ | Android |
| **5** | Twitter for iPad | 1.548.852 | ✓ | iOS |
| **6** | Twitter for BlackBerry® | 1.402.268 | ✓ | Blackberry |
| **7** | TweetDeck | 1.332.883 | | |
| **8** | Mobile Web (M2) | 712.852 | ✓ | |
| **9** | Twitter for Android Tablets | 685.380 | ✓ | Android |
| **10** | Twitter for Windows Phone | 560.787 | ✓ | |
| **11** | iOS | 435.964 | ✓ | iOS |
| **12** | Tweetbot for iOS | 294.736 | ✓ | iOS |
| **13** | TweetCaster for Android | 276.634 | ✓ | Android |
| **14** | Hootsuite | 254.719 | | |
| **15** | Twitter for BlackBerry | 185.251 | ✓ | Blackberry |
| **16** | Mobile Web (M5) | 141.218 | ✓ | |
| **17** | Facebook | 129.358 | | |
| **18** | Instagram | 121.092 | | |
| **19** | Echofon | 116.869 | | |

|  |  |  |  |  |
|---|---|---|---|---|
| **20** | Twitter for Mac | 106.432 |  |  |
| **21** | UberSocial for BlackBerry | 85.506 | ✓ | Blackberry |
| **22** | web | 84.511 |  |  |
| **23** | TweetAdder v4 | 77.204 |  |  |
| **24** | Twitter for Nokia S40 | 75.790 | ✓ |  |
| **25** | RoundTeam | 73.301 |  |  |
| **26** | Plume for Android | 69.811 | ✓ | Android |
| **27** | Twitter for Websites | 61.778 |  |  |
| **28** | twittter-ret app | 59.016 |  |  |
| **29** | AquiHaySeleccion | 39.916 |  |  |
| **30** | http://retweet7ir.com/ | 38.695 |  |  |
| **31** | Write Longer | 34.874 |  |  |
| **32** | dlvr.it | 34.365 |  |  |
| **33** | UsSoccer Responder | 32.033 |  |  |
| **34** | twicca | 31.708 |  |  |
| **35** | BBotMaker - Bot à mots-clés | 30.855 |  |  |
| **36** | Cloudhopper | 29.154 |  |  |
| **37** | Twitter for Windows | 25.780 |  |  |
| **38** | UberSocial for Android | 25.488 | ✓ | Android |
| **39** | Janetter | 23.862 |  |  |
| **40** | Samsung Mobile | 20.180 | ✓ |  |
| **41** | Tweetbot for Mac | 19.150 |  |  |
| **42** | FIFA Mobile Application | 19.025 | ✓ |  |
| **43** | IFTTT | 18.217 |  |  |
| **44** | SPDv2.0 | 16.544 |  |  |
| **45** | والمتابعين التلقائي الريتويت . | 15.477 |  |  |
| **46** | Echofon  Android | 14.766 | ✓ | Android |
| **47** | prueba Ganaseguidores | 14.026 |  |  |
| **48** | Tweedle | 13.161 |  |  |
| **49** | SportsYapper V3 | 12.877 |  |  |

| | | | | |
|---|---|---|---|---|
| **50** | SoloParaDeckApp | 12.468 | | |
| **51** | Tweetlogix | 12.465 | | |
| **52** | Buffer | 12.268 | | |
| **53** | دقيقه لك تلقائي رتويت | 12.229 | | |
| **54** | Windows Phone | 12.119 | ✓ | |
| **55** | TweetCaster for iOS | 11.727 | ✓ | iOS |
| **56** | twitterfeed | 11.702 | | |
| **57** | Talon for Android | 11.680 | ✓ | Android |
| **58** | foursquare | 11.678 | | |
| **59** | Social by Nokia | 11.587 | | |
| **60** | Fenix for Android | 11.508 | ✓ | Android |
| **61** | Seesmic | 11.208 | | |
| **62** | Path | 11.024 | | |
| **63** | YoruFukurou | 10.780 | | |
| **64** | رتوويت32 | 10.729 | | |
| **65** | Vine - Make a Scene | 10.054 | | |
| **66** | Sprout Social | 9.782 | | |
| **67** | Keitai Web | 9.395 | | |
| **68** | aaaas | 8.533 | | |
| **69** | PlayStation®Vita | 8.428 | | |
| **70** | لنك ريتويت توب | 8.246 | | |
| **71** | Botize | 7.730 | | |
| **72** | MetroTwit | 7.688 | | |
| **73** | Carbon for Android | 7.687 | ✓ | Android |
| **74** | Twitterrific | 7.489 | | |
| **75** | ついっぷる | 6.928 | | |
| **76** | UberSocial for iPhone | 6.874 | ✓ | iOS |
| **77** | Twiterous | 6.474 | | |
| **78** | OpenTween | 6.384 | | |
| **79** | Twidere for Android #2 | 6.257 | ✓ | Android |

| | | | | |
|---|---|---|---|---|
| **80** | ShootingStarPro | 5.976 | | |
| **81** | Vine for Android | 5.911 | ✓ | Android |
| **82** | UberSocial Mobile | 5.895 | ✓ | |
| **83** | Клиент для твиттера | 5.872 | | |
| **84** | ツイタマ | 5.830 | | |
| **85** | Tweetian for Symbian | 5.767 | | |
| **86** | Tween | 5.277 | | |
| **87** | Silver Bird | 5.252 | | |
| **88** | ついっぷる for Android | 5.181 | ✓ | Android |
| **89** | jigtwi | 5.121 | | |
| **90** | VidoTimeTwit | 5.013 | | |
| **91** | Gravity! | 4.902 | | |
| **92** | Retwiter صلاح فارس | 4.611 | | |
| **93** | Blaq for BlackBerry® 10 | 4.530 | ✓ | Blackberry |
| **94** | Falcon Pro | 4.240 | | |
| **95** | Dabr | 4.193 | | |
| **96** | OS X | 4.028 | | |
| **97** | TheWorld for iOS | 3.835 | ✓ | iOS |
| **98** | GroupinApp | 3.625 | | |
| **99** | adidasallinarena.com | 3.545 | | |
| **100** | Twitter for Samsung Tablets | 3.492 | | |

# APPENDIX 7: COUNTRY NAMES

The following table displays a list of 194 countries. The second column is the common English name, and the 'Native names' column lists the country's names in the languages spoken in that country.

This list was used to automatically determine Twitter users' locations using their biographies.

| ISO Code | Common name | Native names |
|---|---|---|
| AF | Afghanistan | افغانستان |
| AL | Albania | Shqipëria |
| DZ | Algeria | الجزائر |
| AD | Andorra | |
| AO | Angola | |
| AG | Antigua and Barbuda | Antigua & Barbuda |
| AR | Argentina | |
| AM | Armenia | Հայաստան |
| AU | Australia | |
| AT | Austria | Österreich |
| AZ | Azerbaijan | Azərbaycan |
| BH | Bahrain | البحرين |
| BD | Bangladesh | বাংলাদেশ |
| BB | Barbados | |
| BY | Belarus | Беларусь |
| BE | Belgium | België,Belgique,Belgien |
| BZ | Belize | |
| BJ | Benin | Bénin |
| BT | Bhutan | འབྲུག་ཡུལ་ |
| BO | Bolivia | Wuliwya, Volívia,Buliwya |

| | | |
|---|---|---|
| BA | Bosnia and Herzegovina | Bosnia & Herzegovina, Bosna i Hercegovina |
| BW | Botswana | |
| BR | Brazil | Brasil |
| BN | Brunei | |
| BG | Bulgaria | България |
| BF | Burkina Faso | |
| BI | Burundi | Uburundi |
| KH | Cambodia | កម្ពុជា |
| CM | Cameroon | Cameroun |
| CA | Canada | |
| CV | Cape Verde | Cabo Verde |
| CF | Central African Republic | Centrafrique, Bêafrîka |
| TD | Chad | Tchad, تشاد |
| CL | Chile | |
| CN | China | 中國,中国 |
| CO | Colombia | |
| KM | Comoros | Komori, Comores, جزر القمر |
| CR | Costa Rica | |
| CI | Cote d'Ivoire | |
| HR | Croatia | Hrvatska |
| CU | Cuba | |
| CY | Cyprus | Κύπρος, Kıbrıs |
| CZ | Czech Republic | Czechia, Česko |
| DK | Denmark | Danmark |
| DJ | Djibouti | جيبوتي |
| DM | Dominica | |
| DO | Dominican Republic | República Dominicana |
| TL | East Timor | Timór-Leste |
| EC | Ecuador | |

| | | |
|---|---|---|
| EG | Egypt | مصر |
| SV | El Salvador | |
| GQ | Equatorial Guinea | Guinea Ecuatorial, Guinée équatoriale, Guiné Equatorial |
| ER | Eritrea | ኤርትራ, إرتريا |
| EE | Estonia | Eesti |
| ET | Ethiopia | ኢትዮጵያ |
| FM | Federated States of Micronesia | Micronesia |
| FJ | Fiji | Viti, फिजी |
| FI | Finland | Suomi |
| FR | France | |
| GA | Gabon | |
| GM | Gambia | |
| GE | Georgia | საქართველო |
| DE | Germany | Deutschland |
| GH | Ghana | |
| GR | Greece | Ελλάδα |
| GD | Grenada | |
| GT | Guatemala | |
| GN | Guinea | Guinée |
| GW | Guinea-Bissau | Guiné-Bissau |
| GY | Guyana | |
| HT | Haiti | Haïti, Ayiti |
| HN | Honduras | |
| HU | Hungary | Magyarország |
| IS | Iceland | Ísland |
| IN | India | भारत |
| ID | Indonesia | |
| IR | Iran | ایران |
| IQ | Iraq | العراق |
| IE | Ireland | Éire |

| | | |
|---|---|---|
| IL | Israel | إسرائيل, ישראל |
| IT | Italy | Italia |
| JM | Jamaica | |
| JP | Japan | 日本 |
| JO | Jordan | الأردن |
| KZ | Kazakhstan | Қазақстан |
| KE | Kenya | |
| KI | Kiribati | |
| XK | Kosovo | Kosova, Косово |
| KW | Kuwait | الكويت |
| KG | Kyrgyzstan | Кыргызстан |
| LA | Laos | ປະເທດລາວ |
| LV | Latvia | Latvija |
| LB | Lebanon | لبنان |
| LS | Lesotho | |
| LR | Liberia | |
| LY | Libya | ليبيا |
| LI | Liechtenstein | |
| LT | Lithuania | Lietuva |
| LU | Luxembourg | Luxemburg, Lëtzebuerg |
| MK | Macedonia | Македонија |
| MG | Madagascar | Madagasikara |
| MW | Malawi | Malaŵi |
| MY | Malaysia | |
| ML | Mali | |
| MT | Malta | |
| MH | Marshall Islands | Aelōñin Ṃajeḷ |
| MR | Mauritania | موريتانيا, Mauritanie |
| MU | Mauritius | Maurice |
| MX | Mexico | México, Mēxihco |

| | | |
|---|---|---|
| MD | Moldova | |
| MC | Monaco | |
| MN | Mongolia | Монгол улс |
| ME | Montenegro | Crna Gora, Црна Гора |
| MA | Morocco | المغرب |
| MZ | Mozambique | Moçambique |
| MM | Myanmar | Burma, မ⊞ရန⊞မ◌ာ |
| NA | Namibia | |
| NR | Nauru | Naoero |
| NP | Nepal | नेपाल |
| NL | Netherlands | Nederland |
| NZ | New Zealand | Aotearoa |
| NI | Nicaragua | |
| NE | Niger | |
| NG | Nigeria | |
| KP | North Korea | 조선 |
| NO | Norway | Norge, Noreg |
| OM | Oman | عمان |
| PK | Pakistan | پاکستان |
| PW | Palau | Belau |
| PS | Palestine | فلسطين |
| PA | Panama | Panamá |
| PG | Papua New Guinea | Papua Niugini |
| PY | Paraguay | Paraguái |
| PE | Peru | Perú |
| PH | Philippines | Pilipinas |
| PL | Poland | Polska |
| PT | Portugal | |
| PR | Puerto Rico | |
| QA | Qatar | قطر |

| | | |
|---|---|---|
| RO | Romania | România |
| RU | Russia | Россия |
| RW | Rwanda | |
| KN | Saint Kitts and Nevis | Saint Kitts & Nevis |
| LC | Saint Lucia | |
| VC | Saint Vincent & the Grenadines | |
| WS | Samoa | Sāmoa |
| SM | San Marino | |
| ST | Sao Tome and Principe | Sao Tome & Principe, São Tomé e Príncipe |
| SA | Saudi Arabia | السعودية |
| SN | Senegal | Sénégal |
| RS | Serbia | Србија, Srbija |
| SC | Seychelles | Sesel |
| SL | Sierra Leone | |
| SG | Singapore | 新加坡, Singapura, சிங்கப்பூர் |
| SK | Slovakia | Slovensko |
| SI | Slovenia | Slovenija |
| SB | Solomon Islands | |
| SO | Somalia | Soomaaliya, الصومال |
| ZA | South Africa | Suid-Afrika |
| KR | South Korea | 한국 |
| SS | South Sudan | |
| ES | Spain | España |
| SD | Sudan | السودان |
| SR | Suriname | |
| SZ | Swaziland | eSwatini |
| SE | Sweden | Sverige |
| CH | Switzerland | Schweiz, Suisse, Svizzera, Svizra |
| SY | Syria | سورية |

| | | |
|---|---|---|
| TW | Taiwan | 臺灣, 台湾 |
| TJ | Tajikistan | Тоҷикистон |
| TZ | Tanzania | |
| TH | Thailand | ประเทศไทย |
| BS | The Bahamas | Bahamas |
| TG | Togo | |
| TO | Tonga | |
| TT | Trinidad and Tobago | Trinidad & Tobago |
| TN | Tunisia | تونس |
| TR | Turkey | Türkiye |
| TM | Turkmenistan | Türkmenistan |
| TV | Tuvalu | |
| UG | Uganda | |
| UA | Ukraine | Україна |
| AE | United Arab Emirates | الإمارات العربية المتحدة |
| UK | United Kingdom | |
| US | United States | |
| UY | Uruguay | |
| UZ | Uzbekistan | Oʻzbekiston |
| VU | Vanuatu | |
| VA | Vatican City | Civitas Vaticana |
| VE | Venezuela | |
| VN | Vietnam | Việt Nam |
| YE | Yemen | اليَمَن |
| ZM | Zambia | |
| ZW | Zimbabwe | |

## The detection script

For each country, I executed the following MySQL query (this example concerns Switzerland):

```
UPDATE tweet SET detected_country = CONCAT(detected_country, 'CH')
WHERE `user.location` IS NOT NULL AND
     (`user.location` REGEXP '[[:<:]]Switzerland[[:>:]]' OR
      `user.location` REGEXP '[[:<:]]Schweiz[[:>:]]' OR
      `user.location` REGEXP '[[:<:]]Suisse[[:>:]]' OR
      `user.location` REGEXP '[[:<:]]Svizzera[[:>:]]' OR
      `user.location` REGEXP '[[:<:]]Svizra[[:>:]]'
     );
```

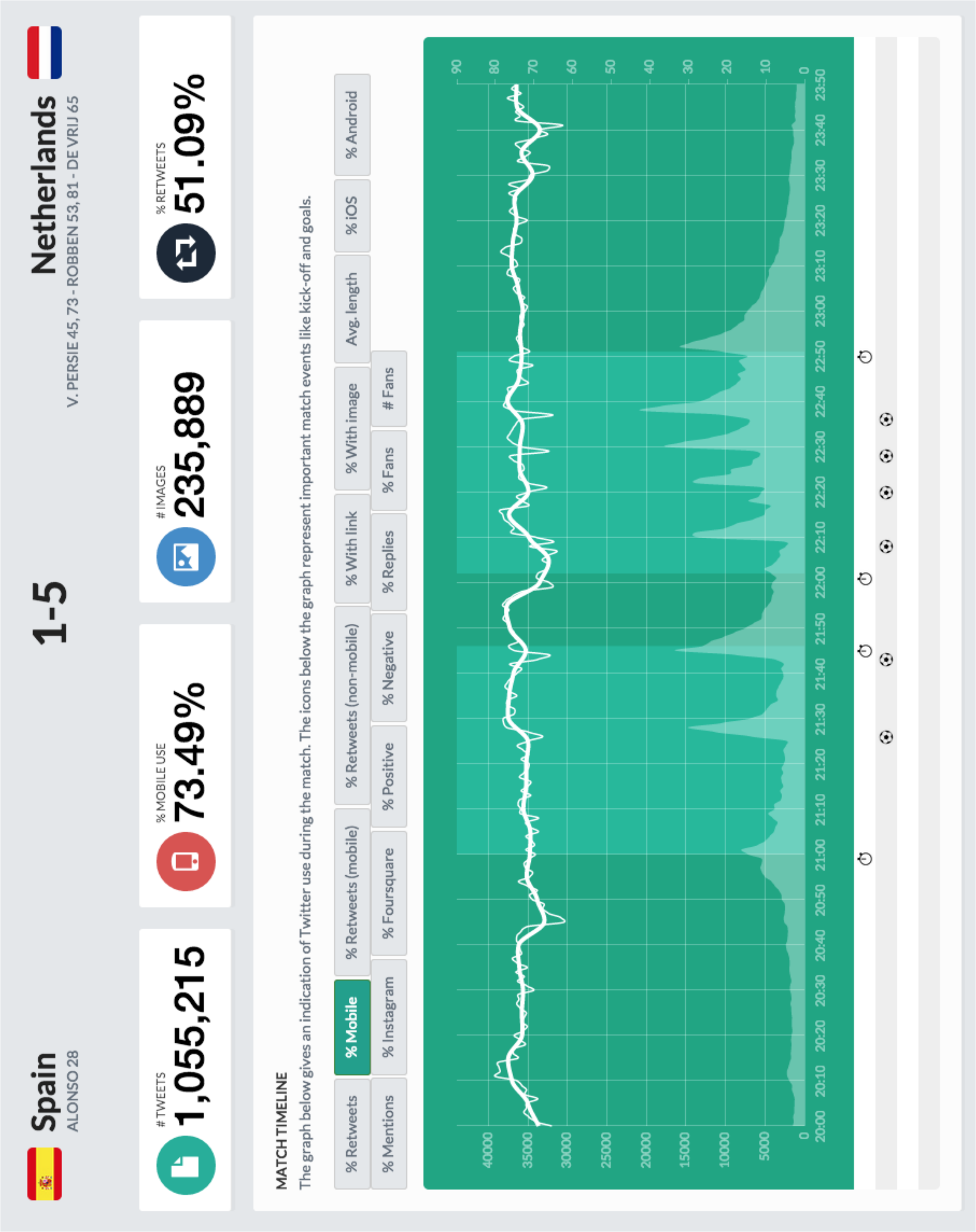
What this query does is to append a country's ISO code to the field 'detected_country' whenever one of the country's names was found in the user biography.

When the above query was completed for each country, I wanted to remove users with multiple nationalities. Of course I have nothing against somebody having that, it was simply impossible to work with in this case. The query I executed was thus:

```
UPDATE tweet SET `detected_country` = NULL
WHERE LENGTH(detected_country) > 2
```

# APPENDIX 8: MATCH TIMELINE

This graphs shows Twitter use during Spain versus The Netherlands. The white line shows the percentage of mobile use, the light-green background rectangles show the first and second half, and the green background area shows the number of tweets per minute.



Spain
ALONSO 28

1-5

Netherlands
V. PERSIE 45, 73 - ROBBEN 53, 81 - DE VRIJ 65

# TWEETS 1,055,215

% MOBILE USE 73.49%

# IMAGES 235,889

% RETWEETS 51.09%

MATCH TIMELINE

The graph below gives an indication of Twitter use during the match. The icons below the graph represent important match events like kick-off and goals.